Vol. 43, No. 6 November, 1946

Psychological Bulletin

THE DETECTION AND TREATMENT OF ACCIDENT-PRONE DRIVERS

H. M. JOHNSON
Tulane University

Most of the material which is to be reviewed here was collected in the period 1936–38, in a study which was being made under the general direction of the writer by the Highway Research Board, National Research Council, for the U. S. Bureau of Public Roads, later reorganized as the U. S. Public Roads Administration. The writer alone is responsible for the accuracy of presentation of the facts and for the proper interpretation of them. The delay in publication is due chiefly to certain problems of national defense which engaged all the writer's spare time from 1939 to 1946.

Modes of Detection

The problem is to segregate a class of drivers of motor-vehicles whose mean accident-rate per unit time or unit distance is significantly greater than that of the remainder of the driver population. The investigators have used two general procedures.

Procedure 1: The examination-method

Direct form. In its direct form, it requires each of the drivers to operate a motor vehicle—preferably over a standard course—under the observation of a trained examiner, who rates his performance. This is essentially a special instance of the "work sampling" method. It suffers from the fact that the sample is not merely small, but is also obtained under conditions that may call into play some unusual attitudes and intentions of the driver. Moreover, it suffers from the fact that it is impracticable to create actual emergencies, which would demand that the driver use his utmost skill, or else wreck the test-car and in so doing expose himself or others to special bodily hazard.

Indirect form. In its indirect form, the examination method requires that each candidate undergo a set of performance-tests of attention, motor skills, information, or the like. Many authors have presupposed that the tasks which they selected require some skills that are "*necessary*

to safe driving"; but certain of them, being more canny than the others, have avoided presuppositions of this kind and have sought merely to build up a battery of tests which give a usefully high multiple-correlation with accident rate. These indirect tests of course cannot reproduce the demands and hazards of actual driving. Some of them, indeed, include situations which are intended to *symbolize* these demands and hazards. But, for example, it is monstrous to presuppose that a subject's response to a *symbolic collision*, in which no glass is broken, no blood let, no quarrels stirred up, and no arrests made, will indicate what the subject might do in an *actual* collision.

Procedure 2: The biographical method

It requires that one accumulate as many data as possible concerning the *personal history* of each operator, including especially his accident-history within a specified period; and attempt to find a set of data obtained by a constant rule, which in their turn will give a usefully high multiple-correlation or association with accident-rate. The principal weakness of this method is that it has not been widely exploited except by certain successful personnel managers who do not publish their results, and ordinary workers have yet to learn to use it to best advantage.*

These two methods are not rivals, but supplementary. Some experts in the selection and classification of personnel have told me privately that they use the examination-method chiefly to "impress" the candidates and their own superior officers; but that in making predictions, they rely principally on the facts that they gather from homely questions about personal histories.

Claims of Some Experts

It is well known that a given set of facts may be useful for actuarial purposes even though they do not validate prediction concerning the performance of any individual. For example, given a set of tests which yield a multiple-correlation of (say) 0.4 with accident-rate, they enable one to pick out one segment of the sample population, comprising (say) one sixth of the total, in which the mean accident-rate is more than twice that of the remainder of the sample; and another segment, comprising another one sixth of the total, in which the mean accident-rate is less than one half the mean rate for the remainder of the total sample; while

* By 1943, a valid procedure was worked out, by means of procedures known as early as 1901. Cf. H. M. Johnson, Multiple Contingency *versus* Multiple Correlation, etc., *Amer. J. Psychol.*, 1944, 57, 49–62, and references cited therein.

the mean rate in the first segment will be more than four times the mean rate in the second segment. To a large employer, such as a department of the United States Government, or a large public utility system, this information can be well used, so long as the applications for employment greatly exceed the number of employees needed; otherwise it would not help.

If insurance companies were organized chiefly for protection of the insured at cost or cost plus a fixed overhead, the same information would enable them to establish differential rates to good advantage (**156**). To an *individual*, seeking personal advice or counsel, however, the same information would be almost worthless, as it would be to a licensing authority or a court, for their duty is not merely to forbid or restrict the use of the highways to those who are likely to create special hazards, but also to see that the others are allowed to use them without annoyance. The two types of problems are utterly different, and much of the confusion which has attended attempts to exploit means of detection for the selection of individuals has grown out of failure on the part of the tester and his client to make the distinction.

Every expert in the field of testing concedes that what we have just said is true of *any* testing procedure, whether actual or ideal. But of those who have based *personal diagnosis* and selective personal treatment on the results of tests that are useful only for *group prediction*, none has yet admitted that he has committed a methodological crime. Some, indeed, have publicly demanded evidence that such nuisances have been committed by anyone. They can find the evidence in these contributions: **2, 3, 4, 22, 23, 24, 26, 37, 38, 39, 42, 43, 45, 47, 49, 53, 56, 57, 59, 62, 70, 72, 74, 79, 81, 82, 86, 87, 89, 90, 91, 94, 97, 100, 102, 104, 106, 107, 109, 110, 113, 114, 115, 116, 117, 118, 119, 121, 127, 1[illegible], 129, 131, 134, 135, 139, 141, 142, 143, 144, 146, 147, 148, 149, 150, 151, 153, 154, 155, 156, 157, 159, 161, 164, 166, 167, 169, 172, 173, 174, 175, 176, 177, 179, 183, 184, 185, 186, 187, 190, 195, 196, 202, 207, 208, 209, 212, 213, 215, 217, 221, 222, 223.**

I now quote enough excerpts to indicate the trend of interpretation that is under criticism.

The official minutes of an open meeting of the National Safety Council's Committee on the Driver, 3 October 1930, record this colloquy:

E. P. GOODRICH (Consulting Engineer, New York City): Do you feel at the present time that you could honestly go to a motor vehicle department and say that it is now advisable to install apparatus in each one of the driver testing stations, and that you can exclude certain drivers from the road on the basis of the tests?

CHAIRMAN BINGHAM: Yes, I would say that, but I would immediately add that I do not think it is as advisable to establish such a department for the primary purpose of excluding drivers as to begin right away to help drivers keep on the road.

Mr. GOODRICH: I would exclude drivers rather than let them practice on the road to the detriment of others.

Dr. Bingham's assertions seem to me to be free of ambiguity, equivocation, and cavil. The tests that were being discussed were those of Whitmer (186); the chairman deemed them suitable for selecting individual drivers for individual treatment.

Again:

The situation is further complicated by the requirement that the tests be accepted as reasonable by both the public . . . and the courts. That is to say, *the tests must obviously measure driving ability.** (Otherwise the applicants could protest that the tests were not valid.) (184, 8).

Specifically, the apparatus has been constructed *to measure certain coordinations involved in automotive driving.* (184, 42)*

This test seems by far the most reliable *for measuring skill in automotive driving* . . . the driver recognizes it as a driving test, and the persons who administer it are not troubled with a lack of cooperation. (184, 50)*

Below are some of the items which can be quite adequately measured and which seem to be *necessary to safe commercial driving.* (112)*

When speed is a factor *the upper 25 per cent only should be selected.* (112)*

It is fundamentally important that the eyes be balanced in acuity. . . . for truck drivers perhaps a good man might get along with 60–70 per cent vision. (112)*

The field of vision must be normal. Restricted field *was found to be highly associated with accident proneness.* (112)†

No one has said the last word on the selection of drivers. . . . the above 13 points will serve as a good starting point and we feel confident that at least 50 per cent of all accidents due to personal factors could be eliminated by rigid application of these principles in the selection of drivers. (112)

The need for good space and distance discrimination is *necessary to safe driving.* . . . use of special stereoscope slides will usually show up this defect. (115)*

Our researches have shown that imbalance of eye muscles is associated with accident proneness. Especially the tendency for the eyes to turn in seems to be detrimental. (115)*

(According to Cobb's report of these tests administered under the direction of the original author, the correlation with accident-rate was −0.21, standard deviation =0.028.)

* Italics mine. H. M. J.

† (In the study reported by Cobb (39) this test, administered under the direction of the original author cited (112) turned out to be correlated with accident-rate thus: $r = -0.27$ for the right eye, and -0.26 for the left eye, each correlation being attended by a standard deviation $=0.21$. Thus, the correlation, though running counter to this interpretation, may be plausibly attributed to chance.)

Various retinal abnormalities such as night-blindness, susceptibility to glare, and similar disturbances *are accident hazards.* **(115)***

(According to Cobb, $r = 0.37$, standard deviation $= 0.18$.)

. . . the objective is to find *which defects are most closely associated with accidents, how much of such defects are necessary* before the driver becomes an accident hazard, and to develop a type of equipment which will make such measurement within the grasp of the average traffic or safety man. **(116, 67)***

Any test . . . must be rigid enough to sample types of behavior or capacities *which will be needed in an emergency.* **(117)***

Optical examinations must be divided into two kinds: (a) those given to *determine whether the applicant is qualified to drive,* and (b) those given to determine his correction *in order to qualify him for driving.* **(117, 1)***

From the results obtained, the driver may be shown a complete picture of his *capacities as a driver.* The tests can be used either *for the selection of drivers* upon employment or for regular checks in order to ascertain the *fitness of the driver at any given time.* . . . **(117, 3)***

Legends on display-cards **(206)**:

"How do you rate as a driver? Find out! Take these . . . driver tests." "Effect of glare. This test measures *how badly* bright headlights affect your eyes." "Selective reaction-time. This test measures *how quickly you* can react in *an emergency.*" "Driveometer. This test measures how well you are able *to manipulate the controls of a car* and to note and obey traffic signs, etc." "Keenness of sight. This test measures how clearly *you can see for driving purposes.*" "Hearing. This test measures *any deafness which may handicap you in hearing warnings.*" "Excitability. This test measures how easily you are excited under unusual conditions."*

The first and most important purpose of this investigation was to analyze *driving ability.* **(43)***

Most of the tests described thus far have been tests for *specific aptitudes that are essential for driving.* **(43)***

The use of scientific driving tests *is the only way of picking out the less fit drivers.* **(58, 5)***

Driving tests . . . are very helpful *in determining whether the state, or fleet owner, is justified in granting the driver the privilege of piloting* valuable cargoes along our congested highways. **(58)***

At last, however, a good beginning is being made on this urgent business of getting up *some real tests of a would-be driver's actual ability* to drive a car safely. **(164)***

Predictive Value of "The Best of All Possible Tests"

We shall presently survey many applications of the two methods which we have just described, and notice the claims to validity for prediction of accident-rate which various exploiters of these tests make. But we shall state first the most important discovery that seems to have been made in the field, and leave it to the reader to evaluate certain other discoveries in terms of it.

* Italics mine. H. M. J.

Cobb (41) presents a long over-due analysis of the problem, pointing out a definitive relationship that is almost obvious in Yule and Kendall (192) and which Greenwood and Yule (85) seem to have overlooked in their important monograph. The number of accidents which any individual has within a given experience, says Cobb, depends on a combination of his personal liability with a set of events which we may call "luck." It is conceivable that the first component can be analyzed, and each of its constituents measured. Should this be done, we could identify *some* constituents as *intrinsic* to the individual's personal make-up. These would include physical and mental deficiencies of various kinds. Others are *extrinsic*, such as the number of hazards created by others whom he meets in the pursuit of his vocation. These hazards are personal in the sense that the specific demands of anyone's vocation is personal. The other component is independent of the individual—it may be compared to the run of cards which are dealt to a player in a fair game. Were the second component ineffective, an ideal test would enable one to predict the accident-rate of each individual exactly. But since chance is always present, the best of all possible tests can do no more than sort the individuals into liability-classes, *such that* the liability is the same for every individual within a specific class, but differing from one test-class to another. But if this were done, the distribution of accidents per operator within each liability-class would follow Poisson's law (41).

Then the variance of the individual number of accidents from A the mean number in each liability-class would be equal to A itself; the variance within the whole distribution from its mean A would be equal to A. Let σ^2 be the total variance from the mean of the distribution; it can be analyzed into two components, namely A and (σ^2-A). A represents the resultant of all the "chance" factors; (σ^2-A) is the resultant of all the systematic factors. It follows that if R is the coefficient of correlation between accident-rate and the scores on the "best of all possible tests," then $R=\sqrt{1-A/\sigma^2}$. Thus given any distribution of accidents per operator, one can determine in (say) twenty minutes what is the predictive value of the best of all possible tests, and decide whether it is worth one's investment of time to try to devise a practicable test. By applying this rule to the Connecticut population which the reviewer studied (40), Cobb showed that the greatest possible correlation R (whether linear or non-linear, simple or multiple) is about 0.46. But among Slocombe's insured operators, drawn selectively from the same parent population, I find that according to Cobb's criterion the maximum obtainable value of $R=0.84$. There is a great difference

in possible predictivity in the two samples. It may be due to more faithful reporting in the insured group than in the general group. For the sample of Boston railway and bus-line drivers described by Slocombe and Brakeman (**157**), the best of all possible tests would yield $R = 0.90$. Although Lauer, in a discussion of my paper (**97**), predicted that no combination of tests would ever yield a high correlation with accident-rate, it is now evident that the outlook is much more hopeful than Lauer then supposed it to be. Evidently, the workers hitherto have made an unfortunate selection of tests; and, since low correlations seem to have been obtained in every instance, it seems likely that the tests which have been used are not of the best possible *type.**

Cobb mentioned also that the value of R necessarily increases with the time over which the accident-rate is determined, according to a definite law.

We do not intend, in this review, to mention each contribution individually and describe its contents. That is the task of an abstractor. We aim rather to point out and discuss some of the most important and definite *trends* that are manifest in the literature.

First: Certain well known investigators have claimed special usefulness for their methods of detection, and their claims have been accepted uncritically by many reviewers. In very few instances, indeed, has any evidence been submitted of the statistical reliability of these procedures; in still fewer has their validity been meaningfully discussed.

For example, we may consider the often cited works of Lahy (**109, 110**) in the selection of operators of motor-vehicles for the transportation-system of Paris. This system is often mentioned as the STCRP—i.e., *Sociêtiê transports en commun de la règion parisienne.* In a single year, Lahy subjected as many as 4,500 operators to an elaborate battery of sensory, attentional, and psycho-motor tests, intended to measure their ability to operate motor-vehicles. The data were at hand when he wrote his book (**109**). Nevertheless, for illustration, he presents the results obtained on only 35 apprentice-operators, and he badly interprets them.

Of these 35 apprentices, the company characterized 30 as succeeders and 5 as failers. Lahy's task was to identify them by means of his tests. This procedure is simple, but we shall describe it in detail. Using Yule and Kendall's notation, slightly modified, let A denote *predicted to succeed;* A' denoting *predicted to fail.* Let B denote *succeeded*, while B' denotes *failed.* Let the bracketed symbols denote *the number of* individuals who have the traits indicated by the symbols. Thus $(AB) = 27$

* It is interesting that Greenwood and Woods (**84**) used Charlier's *Coefficient of Disturbancy* ρ without taking advantage of the relation $R^2/\rho^2 = \sigma^2/A$.

means that there were 27 individuals whom the test predicted to succeed in the trait, and who also succeeded therein. If (B) is fixed, then to maximize correlation, one has to set $(A)=(B)$.* In presenting the results of Lahy's census, the numbers that we set in roman type denote the individuals who were *counted* in each class; the corresponding numbers set in *italics* indicate to the nearest integer how many individuals *would be in* the same class *if* the outcome were independent of prediction.

These numbers are: $(AB)=27$, *26;* $(A'B)=3$, *4;* $(AB')=3$, *4;* $(A'B')=2$, *1*. Thus the number of individuals whose outcome agreed with prediction is $(AB)+(A'B')=27+2=29$. For a perfectly worthless test the corresponding numbers would have been $(AB)_o+(A'B')_o=26+1=27$. Thus the tests properly reclassified $29-27=2$ individuals who would have been improperly classified by chance. The relative net gain therefore is $\Gamma/N=2/35=0.06$ or 6 per cent.

One may express the gain in another manner. The number of individuals who were malclassified by the test is obviously $(A'B)+(AB')=3+3=6$. The corresponding numbers implied by the independency-hypothesis are $(A'B)_o+(AB')_o=4+4=8$ (to the nearest integer). Thus by using the test one would correct $8-6=2$ of the mistakes in classification that one most probably would make by using a perfectly worthless test. This is obviously $2/8=0.25$ or 25 per cent of the total number of mistakes. It might be proper to call the relative saving in mistakes a coefficient of "correctivity," to be denoted by the symbol r'. This coefficient is a partial function of the Pearsonian coefficient of correlation r derived from a 2×2 contingency-table. Based on rounded-off numbers, in this instance, $r'=r=0.25$.

The two relations Γ/N and r' enable one to evaluate the test according to one's purpose. If these numbers are typical of Lahy's whole population, then one must judge that the tests are not extraordinarily effective either as selectors or as correctors of guesses based on the independency-hypothesis and the laws of chance. However, by using an illicit statistical procedure, Lahy and some of his followers such as Husson (**90**) have made these tests look wonderful. Many authors have quoted the coefficients which he reports, along with his interpretation of them. None of them examined the reasoning-processes which gave rise to the interpretation. Had they done so they might have prevented the rise of a psychotechnological tradition that is at least inconvenient.

We ordinarily call a test worthless *if it makes no discrimination;* i.e., if the proportion of malclassified individuals and the proportion of properly classified individuals satisfy the implications of the null-hypothesis H_o, according to which the outcome of training is independent of the test-based prediction. Lahy, however, proposes a different criterion. He proposes to call a test futile if and only if it prognoses that

* This detail is important to some questions to be discussed a little later.

all the test-failers will satisfy the employer's standards of acceptance. In this instance, the only distribution which satisfies this criterion and preserves the actual values of (A), (B), gives $(AB)=25$, $(A'B)=5$, $(AB')=5$, $(A'B')=0$. Thus the agreement between a worthless test and the outcome would be $(AB)+(A'B')=25+0=25$, which is 71.4 per cent of the 35 individuals in the sample. But the number of individuals who are *properly* classified by Lahy's actual test is (as we have seen) $(AB)+(A'B')=27+2=29$, which is 82.9 per cent of $N=35$, or the total number of individuals. Therefore, the difference between the proportion of correct classifications according to this test and according to a test which Lahy calls *worthless* is $82.9-71.4=11.5$ per cent, which is about twice as large as the relative net gain Γ/N which one derives from the definition that I proposed. Lahy's criterion therefore makes his test "look better."

There is good reason, however, for rejecting Lahy's definition of a futile test. For by this definition it is necessary and sufficient that every individual who satisfied the company's criterion had been predicted by the test to fail. Therefore, if the tester should predict that *every* candidate would fail the company's requirements, it would be not only a futile test by Lahy's definition but it would also be the perfect selector in practice, for it detected all the failers. If it predicted that every candidate would *satisfy* the company's parameters it would again be a perfect test in the contra-Lahy sense, that it picked out all the succeeders. Of course, the definition is clear, but one could not use a test that accomplished all this and no more.

Second: Except in the instances which I have mentioned, none of the authors presented a complete description of large samples of drivers, or gave a set of fundamental sub-classes, from which one could reconstruct the original distributions.

It may be that the sample that we have just discussed is actually unfavorable to the results of the tests. It is indeed a poor sample, being too small for proper illustration of the outcome, and having other faults. But it is the best that we have. We might, indeed, consider a sample of switchmen, instead of drivers, who were selected by Lahy's tests. According to Bacquerisse (**3**) and Husson (**90**), $N=198$; $(AB)=76$; $(A'B)=16$; $(AB')=24$; $(A'B')=82$. These numbers enable us to drive $r=0.61$, $r'=0.60$, $\Gamma/N=0.305$, $Q=0.884$. This magnitude of Q is not far off certain other results of Lahy's, but, of course, Q does not, in general, bear a constant relation to any of the parameters in which we are interested.

This paucity of essential information characterizes the reports of nearly all the British and American authors and all the European authors included in this survey who have concerned themselves with

selection of drivers by means of tests. Slocombe and Brakeman (157), for example, present correlations between test-prediction and outcome derived not from their whole sample of drivers, but from a sub-sample that included only the best and the worst. It would be misleading if we considered such a relation as a satisfactory approximation to what we would get if we used all the data. Perhaps, we ought to suggest that although malinterpretations of most of these test-results are still current, they spring from faulty statistical procedures that most of the younger psychologists can now appraise. But these reports are now classical. Unless they are carefully analyzed, they may mislead several more academic generations.

Another study which has been inaccurately appraised is one made by Viteles (177) on a sample of 85 motormen. Let A mean scoring fewer than 4 errors in Viteles's performance-tests. Let B mean safety-record satisfactory to employer. By scaling the author's graphs one can derive the sub-class frequencies that yield the most useful predictions. These are $(AB)=27$, $(A'B)=19$, $(AB')=12$, $(A'B')=27$. Thus we get $r=0.33$, $r'=0.31$, $\Gamma/N=14/85=0.16$, $\chi^2=9.36$, $P=(10)^{-3}$, $Q=0.52$. Note that $Q/r=1.56$. This set of tests has been eulogized by safety-experts and others. It is obviously a little more selective than Lahy's; it could be usefully applied in the construction of an actuarial table; better selectors have long been available; it is practically worthless as a means of predicting the safety-classification of any individual, as such.

Biographical Procedures

The biographical method requires us to collect information concerning the history of every subject in respect to every detail that may seem to be important; and then to classify him first in respect to each detail, and then in respect to his accident-rate. This procedure gives us a set of contingency-tables, but if it yielded nothing further, its value would be small. However, we now have means of treating multiple or joint contingencies, and they are quite practicable. Multiple *contingency* seems to promise more than any other mode of treatment; multiple *correlation* provides the best answer that has yet been found, but it presupposes certain propositions which are psychologically implausible, and to the present writer, it does not seem to have exhausted the possibilities of the biographical method.

Until 1941, the biographical method seems never to have been adequately applied. Among automobile operators, for example, the licensing authorities have collected little information, and they have not yet studied the little which they have gathered. Some large-scale employers have collected a good deal of biographical information, but they have either neglected it or kept it to themselves. But enough has been

done to justify us in trying to discover the utmost limitations of the biographical method and to exploit it as far as we can. For example, in one occupation which we need not name, but which is analogous to all driver vocations, it has been readily and certainly established that the operator's religion, the type and degree of his education, his major interests in school, his previous occupations, his age, the reasons which he mentions for desiring to be trained for the new occupation, and the like, are conjointly associated with the probability of his success in training for the vocation. If some one only devises a proper and adequate method for combining such items of information as these, some astonishingly useful results may grow out of it.

Thus far the only items of biographical information pertaining to accident-liability which have received much public mention are (1) the subject's age, (2) his own estimate of his annual mileage, and (3) his accident-rate within a specified period. Ordinarily, it is not difficult to determine (1), but (2) is very uncertain because very few operators keep track of their private mileage. Whether (3) is a useful item or not depends in large part on the thoroughness with which accident-histories are kept. If an operator becomes involved in a *fatal* accident, the chances are about 9 to 1 that the accident will be recorded; if the accident is non-fatal, but involves injury to some person, the chances are considerably less; and in such a state as Connecticut, where the laws are very strict, the chances of a non-fatal, non-personal accident's being reported are probably even less than that (**174**).

The *accident histories* of a good many public utility companies can be summarized in some such fashion as this: given the average number of accidents which accrue to any specially selected group of employees within this period taken relatively to the accidents of the same operators in an earlier or later period, one may count on getting a coefficient of regression of something like 0.3 to 0.5. This finding is empirical. It holds good for the sample populations which I studied, but if we apply it to untested samples, we should remember that the application depends on some untested assumptions.

To illustrate a type of problem which one encounters in trying to apply the biographical method, I refer to my treatment (**100**, 448) of the study of Slocombe and Brakeman's report (**157**) of a distribution of 7,197 accidents which accrued in 1927 to 2,300 operators of motor-vehicles belonging to the Boston transportation system. In calculating the number of operators in each accident-class according to Poisson's law, I rounded off each number to the nearest integer; thus, the values which I there present (**211**) are not perfectly exact. But, some 96

operators had more than 8 accidents each during the period covered by the census. Chance allows only 19 operators in this accident group. Of the 96 operators included in it, 10 were merely unlucky, while 86 were accident-prone. There is no way of telling which operators belong among the 10 and which belong among the 86. But if we should choose to treat them all alike, the odds are almost 9 to 1 that any specified operator will be properly classified. If our problem is that of the hard-hearted employer, we should treat these 96 high-accident men as accident-prone even though we thereby misjudge 10 of them. If our problem is that of doing justice to the largest possible proportion of individual operators, this treatment would not serve.

Let us notice, however, what would most probably happen if we should treat these 96 multiple-accident makers alike: i.e., if we should eliminate them, or by skillful "dry-nursing" make them over into accident-free operators.

These 96 operators constituted only 4.1 per cent of the whole sample of 2,300 operators, but they drew 1,038 accidents or 14.4 per cent of the accidents which accrued to the whole group. Thus their accident-rate is about 3.5 times the rate of the whole sample. If we could have detected them at the beginning of the year, and replaced them with unselected operators, or operators who had been selected in the same manner as the whole sample, the new set of operators would most probably have had just 300 accidents instead of 1,038. Thus we should have saved 738 accidents or 10.2 per cent of the employing company's total. But there was no way of identifying the operators beforehand, nor yet a way of distinguishing the 10 unlucky normals from the 86 highly susceptibles. Suppose, however, that we begin after the experience has been established. Consider two modes of treatment: namely, (a) do nothing, and (b) eliminate or try to "re-educate" all the 96 trouble-making operators of the year.

(a) I have mentioned that operators who are accident-repeaters in one period tend to regress toward the average of the group in another period, (**98**, **99**). In other words, in the samples which I have personally studied, there is low though positive correlation between the accident-rates of the same operators in two different epochs. Usually its value lies between 0.3 and 0.4. If in this instance its value is near 0.30 and if the two arrays have the same means and the same variances, then procedure (a) would most probably be rewarded by a decrease in the number of accidents within this group for next year; i.e., since this year's number is 1,038, next year's number will most probably be 529. Thus we shall have apparently reduced the accidents of this worst group by

1038 − 529 = 509, or by 49 per cent, just by leaving well enough alone. (Of course, in the meantime, some of the better groups of operators will have been growing *worse*, and therefore may overcome this saving.)

Now let us try procedure (b). Let us suppose that we should "re-educate" the 96 trouble-makers of this year, and thus bring their mean accident-rate for next year to the mean rate for the whole sample of this year. This is 3.129 . . . accidents per operator, which corresponds to 300 accidents among the 96 drivers. Thus we should have saved 1,038 − 300 = 738 accidents or 71 per cent of the total number of accidents which otherwise would have accrued to this group of accident-repeaters. But as we have already seen, treatment (a), which consists in doing nothing at all, would probably have eliminated 1,038 − 529 = 509 accidents, or 49 per cent of all which befell the members of this group. Hence the *net saving* which we can properly attribute to the educational procedure is 738 − 509 = 229 accidents, and this is about 22 per cent of the 1,038 accidents belonging to this group, or 3.2 per cent of the 7,197 accidents accruing to the whole sample.

This net gain may look small, but it is not to be disparaged. If the saving could be extended upon a national scale, it would amount to about 1,000 lives per annum. But, as Crum (**42**) has pointed out in another connection, those who are out to *minimize* accident-rates cannot hope to do it merely by working on a large group of operators who, during a *brief* period, have had an unusually large number of accidents.*

From what we have just said it follows that if we are to gain very much from improving "bad" drivers (**57**), we must not content ourselves with making them over into "average" drivers; we must make them into superior drivers. There are many who affirm that this has been done, but nearly all their statistics are compilations of dramatic instances.

I have discussed this question in some detail because it has been debated emotionally and therefore unprofitably. I once showed (**94**) that in a certain public utility organization, a policy of weeding out the very worst accident-repeaters *quartered* the accidents per mile of operation. This indeed happened. But, from what we have just mentioned, it seems unlikely that the outcome was wholly direct. Possibly the change in policy, which extended into the worst of the years of the depression, had something to do with general morale. Incidentally, in 1936, the company which had the greatest improvement from these purges resumed the policy of the parent company of trying to "salvage"

* In the Boston Elevated's experience, an accident was defined as any mishap which required a report to protect the company against a possible complaint.

its accident-repeaters. In about three years, most of its gains over the other units in the operating system had vanished.

The Government of Trinidad and Tobago (B.W.I.) decided in 1936 on a similar policy of purging its licensed general drivers, (**216**, 1). Because its driver-population is small and its system of fact-finding is extraordinarily good, the effect should be extremely interesting.

Johnson and Cobb (**99**) noted that in estimating the effects of "re-education," especially if the estimate was based on the results of examining the driver for psychomotor deficiencies and the like, some authors had neglected to correct their estimates for a tendency toward regression toward the average, independently of any treatment that might be applied. (**47**), (**49**), (**142**). Later, in a reply to DeSilva (**55**, 1), Johnson showed (**98**) that in another sample population (**95**), (**216**), there occurred, in some accident-making classes of operators, a spontaneous improvement amounting to as much as DeSilva had claimed for his clinically-based advice.

This regression is what Jung (**104**) called "centripetal drift." It often escapes consideration of those who try to evaluate therapeutic agents; it has been strangely disregarded among re-educators of drivers. Examples may be found in references (**5, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19**), each of which deals with one or two individual drivers who were apparently hopeless trouble-makers, but who, for at least a short time after they had been re-educated, or their daily régime had been modified according to advice, remained accident-free. This exemplifies the fallacy of the "dramatic instance" upon which many cures for many ills have been validated to the satisfaction of their inventors. Other examples offer themselves in references (**21, 23, 24, 26, 43, 45, 47, 48, 49, 50, 197**), some of which contain internal contradictions, resulting from interchanged headings and mislabeled scales, that the reviewer was not able to eliminate by personal correspondence; also (**52, 54, 57, 59, 63, 221**), in which it is said that "repeaters are the main root of the autmobile accident problem. . . . The time has now arrived to use the information available, to begin an effective remedial program." Also, (**142**), according to which "most repeaters can be cured of their accident habits." Other examples are to be found in references (**159, 186, 187, 190**).

A striking exception is found, however, in a report (**225**) published by the Personnel Research Federation and attributed to W. V. Bingham, according to which 551 operators belonging to the same unit of a transportation system had collectively 1,670 accidents in 1927, or 3.03 accidents per operator. They were subjected to special personal treatment, which the author does not describe; and in 1929 they had col-

lectively 1,385 accidents, or 2.51 accidents per operator. The saving in accidents was therefore 1,670 − 1,385 = 285 accidents, or 17.1 per cent. This was accomplished by 34 operators being virtually transferred from the class which had 6 to 12 or more accidents per operator to the class which averaged 1 accident per operator.* The probability of the displacement occurring by chance is of the order of $(10)^{-24}$. On the bus lines of the same system the accidents per operator declined from 2.13 to 1.58 in the year 1929, notwithstanding the employment of 192 new operators. The author does not indicate whether these were selected by the same procedures as were the older operators or not. He attributes the improvement to "the detailed studies which have made training and supervision more nearly 100 per cent effective."

It should be noticed that in this example all the operators were worked on—not a sub-class selected according to accident-rate. The employment of 192 new operators suggests that a little purgation may have occurred but the authors say nothing about that.

Age and Accident-Rate

Besides accident-histories in one epoch as indicators of accident-rates in another, we have some important information about the *ages* of the drivers. Until recently, most of our statistics on general accidents were based on the age-classifications recommended by the National Safety Council, which grouped together all drivers younger than 21 years, and grouped the older drivers in 10-year age-classes. A preliminary survey of the writer's (211) suggested that these categories might be too broad to show certain important distinctions. With the generous cooperation of the Department of Motor Vehicles of Connecticut, of Mr. Wm. M. Greene, and of the Works Progress Administration, I procured an age-census, by intervals of one year, of the licensed drivers of the state for 1929 and also for the years 1932–36. The published results (97), (211) disclosed some very significant tendencies hitherto unsuspected. First of all, the fatal accidents per operator within the epoch 1932–1936 were quite unevenly distributed among the age-classes of the operators. For example, the rate among the drivers between 16 and 20 years was 1.73 times the rate for the whole population; 1.93 times the rate of their elders; 2.83 times the rate for the age-group 51–55 years, which has the lowest rate of all the groups that are formed by the rule employed.

Considering the age-classes by single years, one finds that the fatal

* These group-averages require correction for broadness of categories, but I do not have the necessary information.

accident-rate of the drivers 18 years old was only 72 per cent of that of the drivers 16 years old, and only 65 per cent of that of the drivers 19, 20, 21 years old. No satisfactory explanation has yet been offered, but the differences cannot be attributed to chance. The worst record was made by the drivers 19–21 years old, whose rate of fatal accidents per operator per annum was 1.9 times that of the whole population, 2.04 times that of the remainder of the population, and 3.1 times that of the drivers 51–55 years old. The probability of the discrepancy between the rates of the drivers 19–21 years old and the drivers 51–55 years old being due to chance is of the order of 10^{-14400}, the exponent being uncertain in the last two digits. Thus the difference cannot be attributed to chance. Marsh (**123**) sought to relate these age-discrepancies to distances driven per annum, but had to depend on the off-hand estimates of operators. If these estimates are put into the denominator as factors, they make the fatal accident-rate of the worst group, which is composed of youngsters, about 11 times as great as that of the best group, which is made up of middle-aged drivers. Their ratios, however, are at least as uncertain as the drivers' estimates of their mileage, and this reviewer does not regard them as being accurate in the first significant digit.

The non-fatal accidents involving personal injuries and the non-personal accidents which were reported in this study distributed themselves according to the ages of the involved drivers in about the same manner, but both distributions showed less disparity against the youthful drivers than did the distributions of fatal accidents. But the total picture suggests that if a youthful driver is involved in an accident, the accident is more likely to involve death or personal injury than if the driver is older. This finding, which is statistically reliable, may be connected with the finding that the rate of suspension for speeding in the District of Columbia in 1936 is about 2.3 times as high for drivers younger than 23 years as it is for the remainder of the population, and about 11 times as high as it is for drivers 50–55 years old.

Thus, it has been proved that there exist two groups of high-accident drivers: namely, (1) those who are now young, and (2) those who have had an unusually high accident-rate in the past.

The *sex* of the operator establishes a bias also. Women-drivers are *involved in* fewer accidents per operator than are men-drivers. The disparity is not attributable to chance, but it has not yet been shown to depend on sex alone. It remains to determine the influence of such variables *associated with sex* as distances traveled per annum, times of day during which the driver operates the car, nature of traffic and hence

frequency of unusual hazards through which the driver operates, etc., etc. If these considerations are taken into account, it might turn out that sex, as anatomically defined, has little to do with the driver's personal liability. Unfortunately the necessary data have not been collected. Viteles and Gardner (**183**) have presented some convincing evidence that women taxicab-drivers cause many accidents in which they are not legally involved (**182**), by compelling another driver to make an unsuccessful maneuver which may involve a third driver. On the face of the record (**183**), the women show up astonishingly well, but one might do better to look beneath the surface before he formulates a decision.

Biographical data other than age, mileage-rate, and sex, have not in general been exploited in published articles. And yet, Cobb (**40**) showed that a battery of nine pencil-and-paper tests, most of which demanded biographical information only, gave a *shrunken* multiple correlation with accident-rate represented by $R=0.31$ as compared with $R=0.35$ when 22 tests, some involving elaborate instrumentation, were used. Thus the nine items of biographical information removed more than three fourths of the variance removed by the whole battery, and also about 44 per cent of the variance that would be removed by the best of all possible tests. This finding also indicates that biographical information may have been unduly neglected thus far.

Direct Examination

One often re-encounters this hardy recipe: "To learn how a driver operates a car, ride with him and watch him drive." It sounds well, although it does not specify when, under what conditions, and for how long, one needs to watch the driver in order to judge his performance fairly, or in order to make a valid guess of what it is likely to be on some other occasion. We have mentioned some defects that inhere in this procedure. They concern its *selectivity* and also its *adequacy*. About its adequacy, perhaps a few remarks should now be added.

In an ordinary drive, one has a great many opportunities to create an accident. There is at least one such opportunity whenever one meets or passes another car, or passes a stationary object. Let us suppose that the traffic-lane is nine feet wide, that each car which moves in it is six feet wide. If each car keeps to the center of its own lane, it will clear an oncoming car by three feet. If either car deviates from its true course by as much as 5°, it will enter the path of the oncoming car in about 26 feet of travel. This will require about one second if it is traveling 18 miles an hour, and about half a second if it is traveling 36 miles an hour. If the

other car is also deviating from its true course and if it is displaced in the proper direction, these times are correspondingly shortened. Thus, inaccuracies in steering which are quite small in themselves have to be avoided or promptly corrected if collisions are not to occur.

These numerical relations may not be typical of some practical situation which a reader may have in mind. But if he will substitute the numbers which are appropriate to his typical situation, I dare say that he will not greatly change the general picture. To avoid collisions on a modern highway one must steer, brake, and alter one's speed very accurately all the time, or else correct every inaccuracy promptly and quickly. What we have said of actual collisions applies only with greater tolerances to near-accidents also. On a busy highway one may readily encounter 1,000 opportunities for an accident or a near-accident in an hour. If one drives three hours a day on the average, one has on a typical day 3,000 such opportunities. If one drives 333⅓ days in a year at this rate, one has in a year 1,000,000 opportunities. If operator *A* is a very bad driver, he might take advantage of one opportunity in 100,000 and thus have 10 accidents or near-accidents in a year. If operator *B* is a very good driver, he might take advantage of only one opportunity in 10,000,000 and thus have one accident or near-accident in 10 years. Thus, "in the long run," operator *A* should have 100 times as many accidents or near-accidents as operator *B* has. But a "random" work-sample extending through (say) 15 minutes would allow operator *A* only 250 opportunities of demonstrating an affinity for accidents, whereas he needs 400 times as many opportunities as this. Even then, the probability that his rate of accidents or near-accidents within the sampling period is a chance-deviation from zero is nearly 0.50. In respect to operator *B* the chances of catching him in an accident or near-accident in 15 minutes of observation under "random" sampling would be about one in 40,000, and if one did catch him, the probability of his rate within this period being a chance-deviation from zero would most likely be about 0.50. Thus, for neither driver would a "small sample" of his performance enable one to make a reliable prediction about "usual" performance, such as might be afforded by a sampling that extended over (say) ten years.

The difficulty of interpreting the results of a small and "unselected" sample of a driver's performance is multiplied, however, by the *fact* that within a short time, and under unusual conditions, he may assume and retain a set of attitudes, intents, or purposes which he does not assume under the conditions under which he "usually" drives. If, for example, he knows that he is being observed, and that the observer's

report may decide whether he is allowed to use the highways; or if so, when or under what restrictive conditions, he may do his best to impress the observer favorably. Otherwise, he may not care what impressions he makes, or on whom he may make them.*

We probably have said enough to condemn the brief sampling of performance as a basis of prediction of habitual performance in driving, and generally in any other socially valuable activity.† We must now examine the possible merits of the methods called

INDIRECT EXAMINATIONS: I.E., TESTS.

First of all, we ought to consider a certain fallacy of reasoning. It is possible that $r_{xy}=0$, $r_{yz}=0$, $r_{xz}\neq 0$. In one limiting instance it is possible that $r_{xy}=0$, $r_{yz}=0$, while $r_{xz}=1.00$. For example, X, the length of the left arm, is practically uncorrelated with Y, the density of pigmentation of the skin, while Y is practically uncorrelated with Z, the length of the right arm. Nevertheless the correlation r_{xz} between the lengths of the two arms is almost perfect. Similarly, if X is a person's score on a specified test, Y his score in a small work-sample, and Z his accident-rate over a long period, it is possible that r_{xy} and r_{yz} should both be too low to be useful, while r_{xz} is usefully high. Therefore the latter has to be determined empirically, if at all. It cannot be derived from r_{xy} and r_{yz}. One might expect this relationship to be obvious, only it isn't. I

* In one of the United States, it used to be customary to revoke or suspend a driver's license if he were convicted or accused on evidence that seemed to the licensing authority to be probable, of having grossly violated the highway-laws or the traffic-regulations of the state. Before the question of renewing his license could be considered, he had to undergo a new licensing examination. One very intelligent highway policeman, who personally examined one suspended operator, was astonished by his excellent performance during the test. But this astonishment led him to follow the operator for a few miles after he had released him. Within this distance, the operator violated enough laws, regulations, or common-sense rules of safe driving to justify a licensing authority in suspending his license indefinitely, and also in trying to jail him for a very long time.

† It is very hard to convince some persons that they cannot base a reliable prediction of a subject's habitual performance on a very small sample of his behavior. A certain licensing officer once declared: "By the time I have watched a man drive in traffic for a single block I can tell whether he is a safe driver or not." An inspector in one of our aeronautical services once told me: "I can tell by the way a cadet opens and closes his throttle whether he is fit to fly or not." The present reviewer is rather skeptical of such judgments, but even so, he often decides: "Student *A* has what it takes to succeed—he ought certainly to make honors. Student *B* is a mamma's pet—if he passes the course, I shall be pleasantly surprised. Student *C* has a heart of gold and a wooden head; perhaps he can pass but I don't see how." Now many of the reviewer's predictions have failed, as have also those of the licensing officer and the pilot-inspector. But the failures do not seem to diminish the judges' confidence in the accuracy of the predictions which they make later.

have overlooked it; so have many genuine experts in psychometrics and statistics. In neglecting it we have hastily rejected tests that might have been useful in predicting the criterion-trait.

Rival Assumptions about Requirements of Useful Tests

If one examines the procedures employed in various attempts to identify accident-prone drivers by means of tests, one finds three distinguishable *types*, created by three sets of assumptions.

According to the *first* set of assumptions, the test-performance should *imitate* the requirements of the task which is to be evaluated, and the resemblance between test and task should be as close as possible and in as many respects as possible.* It is assumed that it is ideal to make the test-performance equivalent to a sample of the actual performance. If the test-performance is brief, then the work-sample is correspondingly small. Hence, all that we have said against the appropriateness of a small work-sample to the problem applies *a fortiori* to any test which imitates it, for no imitation can be better than perfect.† Thus Weiss *et al.* (**185**), Lauer (**112, 114, 115, 117, 118**), Forbes (**185**), DeSilva (**43, 44**), Lahy (**109, 110**), Tramm (**168, 169, 170**), Miles and Vincent (**129**), Brailovskii (**30, 31**), and others have urged that this resemblance is desirable. Accordingly, one or another of these authors built miniature highways, or devised motion-pictures to be projected upon curved surfaces, so as to create as nearly perfect illusions as they could create, of such problems as arise in actual driving; together with mechanisms actuated by the subject which alter these illusions somewhat as one's manipulation of a vehicle would change an actual traffic-problem. Thus, one experimenter sought to reproduce not only visual stimuli which a genuine railroad track would present to the driver of a moving locomotive, and to make them change as they would change in real life, according to the reactions which the tested driver made to his symbolic controls; he also tried to imitate the rythmic shaking of the platform, the rattling of the trucks over rail-joints, and the sounds of escaping steam (**154**). Thus, by means of motion-pictures, Lahy (**109**) tried to reproduce examples of traffic-situations to which his apprentice-subjects would have to react in their ordinary vocation; thus Forbes in constructing a "miniature highway" tried to make the demands of his test (which involved only *symbolic* steering and *symbolic* acceleration and

* Since *resemblance* is a non-transitive relation, it is not measurable; hence, the fundamental assumptions are invalidated from the beginning.

† As we have seen, it is quite possible to construct a brief indirect test which has a greater predictive value than a work-sample that lasts as long as the test; but this fact seems to have escaped those who hold to these assumptions.

deceleration, and which excluded *symbolic* braking) include as many as possible of the demands which he *presupposed* to be involved in actual driving. Looser imitations than these of the driver's task are reported and defended as being adequate by Lauer (**117**), DeSilva (**43**), Miles and Vincent (**129**), Shushakov (**154**), Wechsler (**184**), and others.

There is a second set of assumptions, however, according to which it is not necessary that the demands of the test should *imitate* the demands of the vocational task. It is sufficient that the test-demands should merely *symbolize* the demands of the task. To the present reviewer, these authors seem to prefer this set of assumptions to set 1: Kafka (**104**), Kehr (**116**), Münsterberg (**135**), Stern (**160, 161**), Sachs (**145**), Schorn (**149**), Roloff (**144**), Moede (**132**), Viteles (**177**).

But there is yet a third set of assumptions, according to which it is not necessary to show that the test-demands should imitate or even symbolize the demands of the vocation according to any specified rule. One merely specifies the demands of the test, and the subject's "degree of success" in meeting them, or the manner in which he satisfies or fails them, together with the corresponding information in respect to the demands of his vocation. One then finds the correlation or association between these paired variables. If the coefficient is high enough, the test is useful; otherwise it is not; but in either event the investigator need not ask *why* the coefficient is what it happens to be. For example, if it should be established that the accident-rates of taxicab-drivers are associated with the number of digits which they can recite in serial order after a single presentation, the reason for the association might be very hard to find; nevertheless one can treat the association as a fact, and make use of it.

Of the investigators whose works we are reviewing, the most canny prefer the third set of assumptions to the other two. One suspects that those who choose either of the other sets may be influenced by their "sales-appeal."

Reliability of Tests

In appraising a test, one needs to know first of all how "reliable" it is: *i.e.*, if the same individuals take it two or more times, how nearly will their scores on the several occasions correspond? Many writers have assumed that the reliability of the test is *proportional* to the coefficient of correlation between the subjects' scores on one occasion and their scores on another. The present writer (**93**) emphasized in 1929 that this reasoning is not only fallacious but also factually misleading; and Bingham (**120**) has repeated the warning. However, if one has given the coefficient of correlation r between the scores of the same sub-

jects in the same test on two different dates, the *correspondence* though not the *agreement* between the two sets of scores is measured, not by r_{12} but by $1-\sqrt{1-r_{12}}$. In other words $r_{11}=1.0$ if $Y=a+b_1 \cdot x$.

Of the authors whom we are considering, most have not given us enough information to enable us to judge the statistical reliability of their tests. We shall mention by name all those who do. We need not concern ourselves about the usefulness of the other procedures until this information becomes available.

Validity of Tests

In appraising a test, we need to know also how closely it measures the independently measurable trait which it is intended to measure. This correspondence is measured by $1-\sqrt{1-r_{xy}^2}$ in which r_{xy} is the correlation between the test and the trait. Only those authors whom we shall mention by name give enough information to enable us to appraise the validity of their tests.

Appraisals of Test-Validities

But many authors whom we cite at least *considered* the duty of appraising the validity of their tests. Some used the rules of appraisal which statisticians accept; certain others invented rules of their own. We shall need to examine the rules of appraisal which were employed, as well as the results which these rules turned out.

We have already mentioned (p. 495) the home made "coefficient de l'accorde" of Lahy, who argued that the results of a test agreed perfectly with the outcome if only the test classified all failers as being failers no matter how large a proportion of the succeeders in the task it happened to classify as failers also. The present reviewer believes that this bit of reasoning ought to be preserved among the curiosa of scientific method and its imperfect imitations. Lahy's interpretation has favorably impressed many commentators, among whom is G. H. Miles (**127**), who speaks of Lahy's test as if it had achieved an ideal toward which other inventors should strive. Other secondary authors have snapped at the same bait.

We have mentioned earlier a test of Viteles (**178**), who presented all the data which were necessary to evaluate it. We mentioned that it correctly classified 12 more operators in a sample of 85 than would be correctly classified by chance *if* the numbers of those who satisfied the employer's safety-standards were what they were and *if also* the number of operators who passed the test (*i.e.*, with a score of fewer than four errors) was what it was. Considered in this light the net gain $\Gamma=12$ op-

erators, which is about 14 per cent of the 85 operators in the sample. Another way of expressing the same facts, however, is in saying that mere chance selection, in these circumstances, would have properly classified 42 operators (instead of 54) and that the excess of 12 operators properly classified by the test should be referred not to the total number $N=85$ individuals who were studied, but to the number $n=42$ who would have been properly classified by chance under the restrictions which we mentioned. Thus, the net gain $\Gamma=12$ operators would be expressed as 29 per cent of 42 operators rather than 14 per cent of 85 operators. Either mode of expression can be defended, but there should be no doubt which mode one employs.

Shellow (**152**) ,who carried on Viteles' work in Milwaukee, compares the records of 166 operators selected by Viteles' tests with those of 163 operators not thus selected. Of the 166 selected operators, 47 operators, 28.3 per cent, were out of service at the end of the first year; while of 163 unselected operators, 65 operators, 39.9 per cent, were also out of service. The probability of the discrepancy being due to chance is of the order of 10^{-2}. Of the 166 selected operators, 33 operators, 19.9 per cent, were lost by resignation as compared with 27 unselected operators, constituting 16.6 per cent of the unselected total. The probability of this discrepancy being due to chance is of the order of 0.2. Of the 166 selected operators, 10, constituting 6.0 per cent, were discharged. Of the 163 unselected operators, 35, constituting 21.5 per cent, were discharged. The probability of this discrepancy being due to chance is of the order of 10^{-5}. Of the 166 selected operators, one individual, 0.6 per cent, was discharged *because of accidents*, while of the 163 unselected operators, 23 operators, 14 per cent, were discharged for the same assigned reason. The probability of this discrepancy being due to chance is of the order of 10^{-3}. Of the 156 selected operators, nine operators, 5.4 per cent, were discharged *not* because of accidents, while 12 unselected operators, 7.4 per cent, were similarly discharged. The probability of this discrepancy being due to chance is of the order of 0.2. Finally, of the 166 selected operators, four operators, 2.4 per cent, changed from the job of motorman to that of conductor, while of the 163 unselected operators, three operators, 1.8 per cent, made this change of vocation. This discrepancy, however, is properly attributable to chance.

Shellow seems not to be describing the same sample-population in every part of her report, for, as the results which we have just mentioned will indicate, the number discharged for "any" cause is not equal to the number of those who were discharged because of accidents plus the number of those who were discharged for other causes than accidents.

Her report contains other items which are inconsistent with the assumption that she is describing a single sample. The reader can find them among the numbers which we have just quoted.

As we have already mentioned Husson (**90**) expressed the results of Lahy's tests of apprentice-drivers in terms of Yule and Kendall's (**192**) coefficient of association Q between the tester's prediction and the employer's decision to hire or not hire the apprentice. As Yule and Kendall have pointed out, there is no necessary relation between the coefficient Q and such coefficients as r, T, C, and ϕ, in which Husson is interested; and the interpretations which he puts upon Q as an approximation to Pearson's r, for example, are quite illicit. For example, Husson mentions one Q-coefficient of the order of 0.88 between test and outcome, which he interprets as if it were an r-coefficient. The writer has obtained particular 2×2 distributions which yielded Q-coefficients greater than 0.8 and r-coefficients of the order of 0.3. One therefore cannot find r-interpretations or χ^2 interpretations for Husson's findings in the information which he gives.

In validation of certain other selective tests, the evidence is about as quaint as Lahy's. For example, in the street transportation system of greater Berlin, Tramm (**170**) compared the accidents which accrued to 50 unselected operators with those which accrued to 50 operators selected by his tests, through 21 months, beginning with the date of employment. At the end of the 21 months' experience the unselected operators had accumulated 31 accidents while the selected operators had accumulated only 22. The difference is not certainly attributable to chance, and not certainly to systematic agencies. Its history, however, is extraordinarily interesting. By the end of the eighth month, the 50 unselected operators had accumulated 21 accidents and the 50 selected operators had accumulated 18 accidents. In the next 13 months the 50 unselected operators accumulated 19 accidents while the 50 selected operators accumulated 9. In some of these intervening epochs, the 50 unselected operators had collected more accidents than the 50 selected operators, but in certain other epochs, fewer. Nevertheless, at the end of 21 months, the 50 unselected operators had accumulated 9 more accidents than the 50 selected operators. Tramm and some of his reviewers have made much of this difference. However, the facts agree with the hypothesis that both classes of operators were drawn by the same sampling-procedure from the same parent-population. We might compare the two classes of operators to two race-horses, and compare their accident-collections to ground covered before a specified instant. Suppose Horse A starts more quickly than Horse B but that eventually Horse B

overtakes him. At every instant meanwhile, Horse *A* will have covered more distance than Horse *B* and will have shown a greater average speed. Now, of these two classes of operators the 50 who were unselected correspond to Horse *A*; the 50 test-selected operators correspond to Horse *B*. At the end of 21 months, these two classes of operators have run about 6 per cent of their course. At this stage, the unselected operators have accumulated more accidents than the selected operators, and have therefore attained a higher accident-rate. But it would be very rash to extrapolate from 6 per cent of the total experience as a means of estimating what the comparison will be in the remaining 94 per cent of the experience. Hence, it might seem wise to discount Tramm's comparison.

Vernon (**174**) after describing a battery of aptitude-tests employed by the (British) National Institute of Industiral Psychology, and now well known in America, presents two validating exhibits. Exhibit *A* shows the order in which 12 subjects arranged themselves according to their total scores in the Institute's tests, and also their order of merit as drivers according to the employer's rating, which, so they assure us, was independent of the tests. The rank-order correlation was 0.94. This, of course, signifies a very close correspondence between the two ordinal ratings. Since the Institute is reputed to have accumulated paired ratings on some thousands of drivers, one might expect it to publish an exhibit based on some large number, especially if such a relationship as this prevailed in a large experience. Many years have passed since this report was published. The additional information has not yet appeared. I offer no interpretation.

As Exhibit *B*, Vernon (**174**) showed the rank-order scores in this battery of tests paired with a ranking based on the accident-records combined, in a manner which he does not specify, with the opinions of their supervisors on matters which he does not mention in detail. The rank-order correlation in this exhibit is about 0.78. Again, one wonders whether this exhibit is typical of a *large* experience which the Institute has had opportunity to accumulate, and why in any event the Institute has not yet published the comparison. Surely the Institute should not expect one to overlook this omission or to consider that its tests have been validated as accident-rate predictors from a sample of 18 subjects selected in any manner whatever from a large population.

Miles and Vincent (**129**) recite a part of this legend; so does Myers (**137**), who accepts and proclaims uncritically the exhibit of Tramm which we have just analyzed. Miles and Vincent mention (**129**) a proposal in Parliament that certain of the Institute's tests be required of

all applicants for drivers' licenses. The Institute's fee was then two guineas, or about $10.00. The Parliamentary Committee reported against the proposal. If the decision was based wholly on such evidence as we have reviewed then it seems to have been well-founded.

I have mentioned this matter in detail because some propagandists for testing programs have suggested that if a test or a program for testing originated abroad, and especially in Germany, France, or England, *therefore* it must be valid. The fact is that this Institute seems to have exploited commercially a set of aptitude-tests for drivers, aviators, and the like, which it had not validated or standardized. Henceforth, one may well examine its offerings with some special care.

In contrast to counterfeit validation, we may examine the report of Cobb (**40**). In 1936 there were "on the market," so to speak, two sets of tests which resembled those of the National Institute, Lahy, or Tramm. The American authors had improved some of them and had made them more attractive to the eye, but speaking roughly, each of them was intended to measure some skill or set of skills that somebody had considered to be "necessary to safe driving" and had recommended them to the consideration of the licensing authorities or the courts. Cobb sought to administer each of these tests to 3,663 drivers* licensed in Connecticut. This state records the general accidents, violations, convictions, etc., with unusual care, and perhaps more nearly completely than any other state does. However, as I have mentioned earlier (**97**, **211**), the records are not complete, except with respect to fatal accidents. It is probable that not more than half the accidents which are reportable under the law, but which do not involve deaths or personal injury are actually reported, unless the drivers are insured.

Cobb's distribution is non-typical of the general driver population of the state, for it contained many drivers whom the authorities had sent in for examination because they had recently got into trouble. This action was defensible on administrative grounds, but it did bias the distribution, and thereby limit the permissible interpretation of the results.

These tests were 72 in number. They were being administered in the field and were also being recommended for use in *personal* diagnosis or advice, some by the department of psychology in Iowa State College and some by the Harvard Bureau for Street Traffic Research.† Their authors having described them elsewhere, as Cobb has also done, we need

* Among the persons who took the tests, some 502 could not be identified in the records of the state. Some of them were visitors from other states; others were under age; while still others were otherwise disqualified.

† Later transferred to Yale.

not enumerate them here. In respect to equipment, some required nothing but pencils and paper and no specially trained personnel; one required $5,000 worth of apparatus and a staff of experts. As we mentioned above, the predictive value of the pencil-and-paper tests alone is almost as great as that of the whole battery; so that it would be far more economical, though less spectacular, to improve them than to retain for any future work the tests which require apparatus that is usually very expensive and requires much service to keep it running.

These tests were administered under the immediate and personal direction of those who had devised and sponsored them: namely, Dr. A. R. Lauer, for one set, and Dr. H. R. DeSilva and Dr. T. W. Forbes for the other. Since the Highway Research Board and the Bureau of Public Roads wished to determine the validity of the tests as selectors of operators according to their susceptibility to accidents, their sponsors were instructed to administer them as they were used to administering them in the field. Although this stipulation was proper, it caused some potentially valuable information to be lost.

We may now consider four especially important hypotheses which were adequately tested in this experiment.

Hypothesis 1: The tests enable one to select classes *of operators, such that every class will include enough operators to make the classification useful, while the accident-rates will differ widely and reliably from one class to another.*

The results verify the hypothesis. Although the correlation between multiple test-score and accident-rate is low, nevertheless, the 18 per cent of the operators who earned the highest scores had less than half the accident-rate of the others, while the 17 per cent of the oprerators who earned the lowest scores had three times the accident-rate of the others. The accident-rate of the latter selected group was about 4.7 times as large as the accident-rate of the former. According to probability-tables these differences are almost perfectly reliable. Thus the tests have a distinct field of usefulness; namely, to large employers and to insurance-companies.

Hypothesis 2: The tests enable one to select individuals *whose accident-rates vary widely and reliably.*

The results invalidate the hypothesis. The correlation between weighted multiple-score and accident-rate is about 0.35, which means that the tests are about 6 per cent as effective as a perfect test would be as a means of predicting accident-rates of *individual* operators. In other words, for this purpose they are not worth the cost of administering them and taking account of the scores.

As Cobb (41) has noted, these subjects were selected, *partly because of their accident-rates* from a parent population which has certain known characteristics. Among them is that if one should devise an ideal test that *perfectly* classified the individual operators according to their *susceptibility* to accidents, it would nevertheless be very feebly correlated with their accident-*rates*. But if the susceptibility and the rate are quite independent of each other, the administrator should care nothing about susceptibility.

Hypothesis 3: That over a long period the authorities might do well to treat the operators selectively, according to their test scores.

The results contradict the hypothesis, in falsifying Hypothesis 2. Aside from the consideration whether remedial or educational treatment of high-accident operators is effective, the results show that even in the 18 per cent of the operators who make the lowest scores, the majority have gone accident-free so far as the records show. It is not to be expected that within a democracy, the authorities could impose selective restrictions or selective treatment on any class of operators, the majority of whom show no need for it.

Hypothesis 4: That the personal weaknesses disclosed by the tests are related to types of accidents to which the operators are especially susceptible.

For example, it has been supposed that if a person is especially sensitive to glare, then he is unusually susceptible to collisions at night; if his judgment of the distances of visible objects is inaccurate, then he is unusually liable to head-rear collisions, etc.

The *test* of the hypothesis consisted in this: Classify the operators according to their scores in the testing trait X, and for every X-class determine what proportion Y of their accidents belong to the type to which they were supposed to be susceptible. Select a critical score such that between the individuals who satisfy or surpass it and the individuals who fall below it, the Y-differences will be as great and also as reliable as possible. Then ascertain whether the Y-differences are important and trustworthy.

The results may be summarized as follows:

1. *Sensitivity to glare*, whether determined by the Harvard method or the Iowa State method, is not significantly associated with preponderance of night accidents over other accidents.

2. *Defects of color vision* are not significantly associated with a preponderance of accidents at signal-protected crossings over other accidents.

3. The Iowa test of *distance judgment* is not significantly associated with a preponderance of head-against-rear accidents over other accident. Neither is the Harvard test.

4. *Inequality of acuity* in the two eyes is not significantly associated with a preponderance of head-against-rear accidents over other accidents, whether the trait is determined by the Iowa State procedure or the Harvard procedure.

5. The time required for simple reaction to auditory stimuli is not significantly associated with propensity to head-against-rear or passing-collisions over other accidents.

6. The *variability* of these reaction-times *is* associated with a propensity to head-against-rear or passing-collisions over other types of accidents. The association is in the expected sense although it is weak. Its strength may be expressed by the coefficient of correlation $r = 0.06 \pm 0.03$, or by the fact that the test properly classifies 28 accidents out of 1,487 that would have been improperly classified by chance. In other words, $\Gamma/N = 1.9$ per cent.

7. The subjects' scores on the Harvard vigilance-test ("sensory braking"), in which the subject was to depress a brake-pedal as quickly as possible in response to a visual stimulus, are not significantly associated with a propensity to head-against-rear or passing-collisions over other types of accident.

8. The scores on the Harvard vigilance-test ("combined braking"), are not significantly associated with propensity to head-against rear or passing-collisions over other types of accident. (In this test the average time between the removal of the subject's foot from the accelerator-pedal and its effective application to the brake-pedal is taken.)

In summary: In only one of these tests did any association between test and predicted trait manifest itself, namely in number 6. In this instance the association is in the expected sense, but the coefficient is too small to justify its use in selecting special advice or specail treatment for any individual. In all other instances the association is non-utilizably small; and in respect to the tests mentioned under headings 1, 2, 7, and 8, *runs counter to the prediction* of the advocates of the tests. Considering these sets of tests in the light of their bases of rationalization, we may say fairly that this basis is not justified in any instance except the one mentioned under heading 6, and that whatever its statistical probability may be, nevertheless it is not sufficiently discriminating to justify its use.

In many analogous situations one has to deal with tests which are advocated not for the validity or predictive value which they have been shown to have, but rather for their so-called face-validity, *i.e.*, the validity which some expert has *presumed* them to have. This exemplifies Münsterberg's procedure, the history of which is not fragrant as orange blossoms and roses ought to be.

Of course, the value of a test for one purpose is independent of its

rationalization for another purpose. Perhaps these tests, though badly rationalized, may find some usefulness for other purposes than those for which their inventors recommended them.

Engineers and inventors of safety-gadgets have had their say in accident-prevention. They have contributed much that is valuable. Psychologists and physiologists have made good suggestions, most of which did not stand the test of utility. Optometrists promised for a while to ease into the picture, but their leaders demonstrated that they had a healthy respect for facts. There are some psychologists who, having learned something of the past, have conceived and even tested some possible procedures of promise, which they are willing to examine more closely when the states take up the accident-problem in earnest. Meanwhile there are psychiatrists who believe that the task should be committed to their able hands, if and when their proposals can be adequately financed.

Notwithstanding this, many psychiatrists have been extraordinarily patriotic. They are quite willing—if their activities can be financed—to decide whether individuals who have passed other elimination-tests should be finally admitted to a vocation or to training for a vocation. They are even willing to aid in reviewing the decisions respecting persons who have been disqualified on other than psychiatric grounds. They argue that in order to understand an individual's behavior, one has to consider him "as a whole"—whatever that may mean. Somehow, they suggest rather than argue that they are especially qualified to consider him in this manner. The fact that in making diagnoses they do not bind themselves by the rules of reasoning that restrict scientific workers is, of course, another advantage. Among these impediments are rules against affirming the consequent, denying the antecedent, equivocation, and disregarding negative instances.

The report of Raphael *et al.* (**143**) exemplifies this general procedure in certain respects. Their sample-population was 100 English-speaking traffic-offenders, selected in order of their apprehension, and referred by the local recorder's court to a psychopathic clinic. Most of them were pronounced to be psychopathic, 1 as being "psychotic or insane," 1 as having an active epileptiform "tendency," at least 3 as being "very seriously handicapped physically," 6 as having "significant defects" in hearing, and 14 in vision; 4 as having abnormal color vision, 46 as being "seriously handicapped by alcoholism," which, I suppose, means that they sometimes drink more than is good for them. Moreover, 48 of these troublemakers were said to be "markedly inadequate and suggestible," 41 as "lacking in alertness," 57 as "emotionally unstable and

impulsive," 26 as "excitable," 14 as having "immaturity of attitude and response," 28 as "unreliable or undependable," 31 as "egocentric," 45 as "lacking a proper sense of responsibility." 3 as being "actually anti-social in attitude and *tendency*," 42 as having "inferior intelligence." The total number of defects noted was 370, an average of 3.7 defects per individual. The total exhibit reminds one of Lewis Carroll's puzzler about the Chelsea Pensioners,* in that it makes one wonder how many *at least* must have had any two, any three, . . . , of these defects. It would be a negligent psychiatrist who could not find *some* mental defect in anyone whom he carefully examined and compared with an *ideal norm*, for none is perfect but God.

However, we may very seriously and earnestly inquire whether this sample is abnormal in a statistical sense. For example, consider color-vision. Is it remarkable to find 4 per cent of *any* sample population defective in color-vision? Statistics vary, but 4 per cent is often given as typical.

Again, consider "inferior intelligence." This trait is operationally defined (so one of these authors courteously assured me in a private letter) by Stanford-Binet IQ, in which 16 years was called the basic age for adults. The standard therefore is a *statistical*, not an *ideal* norm. Would it be unusual to find any group of 100 individuals selected *at random*, of whom fewer than 42 per cent were "inferior"? The correct value ought to be near 50 per cent. But to make sure of the matter, I reduced to a common basis these scores and those of the sample which Yerkes (**189**) submits as typical of the United States Army draft of 1917. I then applied Pearson's χ^2 test for homogeneity, trying to falsify the hypothesis that both samples were drawn by the same selection-procedure from the same parent-population. The result is consistent with the hypothesis. The probability of the discrepancies being jointly due to chance is about 0.70. Who believes that in a democratic state it would be possible to eliminate 42 per cent of the general drivers on the ground of "inferior intelligence?" And especially when one considers that most of them vote and have kinsmen who can vote? And furthermore, when the correlation between intelligence and accident-rate is too low to enable one to make reasonably safe predictions about any individual as such?

I pass over such subjectively defined traits as "excitability," wondering how stolid a person would have to be if he remained placid while he

* "A Tangled Tale, Knot X." In *Logical Nonsense*, the *Works of Lewis Carroll* (C. L. Dodgson), edited by P. C. Blackburn and Lionel White. Putnam's, 1934. Cited by Yule and Kendall (**192**).

was being involuntarily subjected to a psychiatrists' inquisition, or how many of us could clear ourselves of a charge of "egocentricity"or of "lacking in alertness." I wonder also whether fewer than 46 per cent of the psychiatrists of this nation are "seriously handicapped by alcoholism" in the sense in which this term is defined above.

Fortunately 13 of these offenders were diagnosed as being "acceptable drivers on liberal evaluation." That may be about the percentage of ordinary individuals whom the psychiatrists would pass if the decision were left to them. I for one should like to see their *statistical* norms, if they have any, and to learn how they derived them.

Canty (**37**) attempts by psychoanalytic procedures to find what is wrong with traffic violators; he includes, however, "a complete physical and neurological examination including blood serology." He also includes and recommends several mental and motor tests which the studies of the Highway Research Board have demonstrated to be without value in individual diagnosis. Canty's group might have made the same finding if they had considered *non-offenders* along with a sample of those who were caught.

Salmon (**146**) devotes 11 pages to a statement of his clinical impressions as a psychiatrist, which interest us in that they belong to an interesting man. He believes that "there can be no question of the propriety of investigating from this point of view the minds of operators who have participated in accidents"—and why not the others?— "and grouping by fundamental methods of all sciences the causes and effects that are discovered."

Selling's article (**150**) is addressed to his medical brethren, and reproaches them for having "permitted this testing activity to pass beyond their control." Referring to so-called intelligence-tests of eyesight, etc., he says: "Until *physicians themselves give* these examinations, *compile data*, and *show just where* the line *must* be drawn between adequate and inadequate physical capacities, licensing by means of physical and mental tests will be more or less of a farce." (Italics mine.)

The licensing examination as practiced in most of our states is indeed farcical. But one may well doubt if it would become non-farcical if the tests were increased in number, and administered by physicians, who would compile the data, and by interpreting the statistics "show just where" the line could be drawn between enabling and disabling "capacities." For, it can be safely predicted that this "line" cannot be drawn at all. Indeed, as we have already mentioned, a certain medical scientist (**41**) has proved, by sound statistical procedure, that the best of all possible tests can do no more than segregate the individual operators

according to their *liability* to accident. Within these liability-classes the accidents per operator will vary according to the laws of chance, although the rate will vary from one class to another. In other words, the most that can be done is to set up actuarial tables. We cannot safely prognose the performance of any individual as such.

There are, of course, physicians who can devise and administer psychophysical tests. There are physicians who can employ and interpret the appropriate statistical procedures. There are some who can do both. These last are the only physicians who meet the requirements that Selling outlines. They are few in number. Still fewer are they who also are continuously available for such duty.

As to the *administration* of even the simplest psychophysical tests: I have seen analyses prepared by D. R. Brimhall and R. Franzen of the medical records of commercial aviators who were examined periodically and also on special occasions. These results have not yet been published, although certain organizations of physicians are now acquainted with them, and others can easily become so. I suppose that every reader will recognize that a test of visual acuity, or of a similar function is essentially a psychophysical test, in that it implies the determination of a *threshold*. But a threshold is essentially a *statistical* concept. To satisfy its definition the examiner has to shut out many extraneous variables which otherwise would bias the result. As every psychologist knows, the order of presentation of the stimuli affects the result; so does the subject's mode of expression; so does the form of the examiner's questions. Are these precautions known to those physicians who are to make these examinations? To some—yes. They have been trained in experimental psychology. Is this discipline included in medical and premedical curricula? Sometimes yes, usually no. How, then, is a physician to become acquainted with these procedures? Of course, there's much that an intelligent individual can learn for himself if he is well motivated and has the time. But, about the time? Let us suppose that the physician is busy curing the sick, healing wounds, etc., then what?

Now if two physicians use the same examining-procedures on the same individual and if the patient's condition does not greatly change in the mean time, then they ought to get roughly the same results. By this phrase I mean something more than saying that their results should be closely *correlated*, for they could be perfectly correlated and still be infected with a huge constant error associated with the individuality of the examiner.

And yet, analysis of the examiners' own records, such as Brimhall and Franzen made, indicates an almost incredible disagreement on the

simple question whether a given patient has *any* visual deficiency that unfits him for duty. As to psychopathological diagnoses: Of 58 patients examined by two physicians (and not always the same two) and disqualified by at least one, their recommendations agreed on only five cases. Thirty-eight of these candidates were disqualified by the first examiner for defects the second examiner did not find; 15 of them were disqualified by the second examiner for defects that the first examiner did not find.

It is, of course, notorious that in criminal trials or in civil trials of contested wills, in which the *sanity* of a person is questioned, each side can usually find one or more psychiatrists to testify in its favor. But if such wide dissent as this is general, and if it should become generally known, then the courts might find it hard to rationalize the admission of psychiatrists as expert witnesses, and the licensing authorities for employing them as expert consultants.

Again, the duty of compiling these records and interpreting the census requires an expert statistician. There are some such experts within the medical profession. Indeed, within one small division of the Bureau of Medicine and Surgery of our Navy, and before the outbreak of this war, there were several of them, their primary specialties including such diverse fields as cardiovascular behavior, cancer, hospital administration, ophthalmology. The value to the Navy of their understanding of statistical procedures can hardly be overestimated.

But biostatistics is not yet heavily emphasized in most of our medical schools, and most busy practicing physicians lack the time, or the incentive, to train themselves. Of course, it is to be expected that the picture will change within the present generation. Meanwhile most medical statistics, except those put out from institutions that maintain statistical bureaus, are home-made. How bad they are is suggested by a medical scientist who set up the statistics division of the Mayo foundation and clinic.*

* H. A. DUNN, Applications of statistical methods in physiology. *Physiol. Rev.*, 1929, 9, 275–398, examined "200 medical-physiological quantitative papers from current American periodicals." He judged that "in over 90 per cent statistical methods were necessary and not used. . . . "

"In almost 40 per cent conclusions were made which could not have been proved without . . . some adequate statistical control. . . . "

"About half of the papers should never have been published as they stood; either because the number of observations were insufficient or because more statistical analysis was essential" (p. 276).

Of course, the picture has improved since 1929. Nevertheless, most of the physicians who are now prominent were trained before that year.

But, there is nothing in the medical curriculum which can qualify a physician to pass judgment on a statistical question except rigorous training in statistics.

Perhaps in no branch of medicine does one find a greater lack of statistical intelligence or a greater aversion to statistical procedures than in psychiatry. This field has been deeply invaded by psychoanalysts. In fact, it is dominated by them. And, as everyone knows, the Freudians, Jungians, and Adlerians rationalize their practices by an *antilogic* or *paralogic*, in which the fallacies of affirming the consequent, denying the antecedent, equivocating, disregarding negative instances, are treated as if they were formally valid. To any one who uses these processes, the tedious business of sampling, selecting unequivocal criteria for classification, classifying, counting, and determining valid descriptive laws, with indexes of the range of their applicability, is as a sword in one's bones or as poison ivy upon a scratched and sunburned skin.*

I recently saw a letter written by an eminent psychiatrist who had proposed to predict success or failure in training or practice in a certain vocation from his own unanalyzed impressions of *somatic types*. On its face his proposal looked preposterous. But a sample was taken, protected against contamination in the sense that his crew knew nothing about the candidates' training-record and the training-department knew nothing about the judgments of his crew. The correlation between his predictions and the final outcome was not negligible. Perhaps it wouldn't "go" as a full-page article in a Sunday magazine, but nobody could laugh it off. He desired financial support from the organization that had accepted his original experimental design. That organization was willing to spend its money but it favored a more extensive and thoroughgoing test. But he said that the ideas of statistically minded persons were so different from those of "the psychiatrist" that he would rather have no more to do with their procedure or with them.

Thus one may well doubt whether practicing physicians and especially practicing psychiatrists are the best of all qualified persons to monopolize or even handle the procedures that are necessary and sufficient for administering and evaluating tests of aptitude for driving.†

I may add that in predicting success in training for a certain other

* There are, of course, exceptions. One psychiatrist I have known who used and understood the slide-rule. He even became interested in Chi-squared tests of independence and the like. But his military duties proved to be too heavy. In an effort to dodge work, he reverted to type. I seem to remember that he adopted Holism as a means of rationalizing his escape. Still, he escaped.

† A good *medical scientist* may be happy in a salary of $5,000 to $8,000 a year. A clever psychoanalyst can get a *minimum* of $10.00 a half hour chiefly for listening to introspective misfits describing their imaginary problems, fantasies, and memories. Of course, all psychoanalysts are public-spirited. They admit it. But how much time can any of them spare for routine-examination of automobile operators, and at what price? And are enough of them available to solve the accident-problem if they could?

vocation, somewhat akin to driving, the psychiatrists' predictions contributed nothing to the efficiency of a combination of those predictors which satisfied Wherry's criterion for admission to the test-battery. These results were uncontaminated, in that the medical examiner knew nothing of the training-department's records, and vice versa.

On the whole, we may doubt whether Selling has established a valid claim for physicians and especially psychiatrists being granted a monopoly of administering and evaluating testing-procedures. In the long run, the validity of their judgments must be appraised by exactly the same procedures as those of any other group. To them there belongs no priestly exemption. And fortunately many of them do not claim it.

BIBLIOGRAPHY

1. Ach, N. Psychologie und Technik bei Bekampung von Auto-Unfallen. *Industr. Psychotech.*, 1929, **6**, 87–97.
2. Allgaier, E. A portable chronoscope. *Amer. J. Psychol.*, 1935, **47**, 685–688.
3. Bacqueyrisse, L. Psychological tests in Paris tramway and omnibus services. *Human Factor*, 1935, **9**, 231 ff.
4. Baker, J. S. Do traffic accidents happen by chance? *Nat. Safety News*, 1929, **20**, 12–14.
5. Baker, J. S. What can we do for high-accident drivers? *Nat. Safety News*, 1932, **25**, 19 f., 63.
6. Baker, J. S. Finding the high accident drivers. *Public Safety*, Jan., 1933, **7**, 20–23.
7. Baker, J. S. Wise fleet men watch the driver's mileage. *Public Safety*, Feb. 1933, **7**, 22 f.
8. Baker, J. S. Accident clinic: The case of Mr. S. *Public Safety*, Feb. 1934, **8**, 23 f.
9. Baker, J. S. Accident clinic: The high price of relaxation. *Public Safety*, Mar. 1934, **8**, 20 f.
10. Baker, J. S. Accident clinic: Two jobs one too many. *Public Safety*, Apr. 1934, **8**, 20 f.
11. Baker, J. S. Accident clinic: The case of Hugo and John. *Public Safety*, May 1934, **8**, 20 f.
12. Baker, J. S. Accident clinic: One bad habit nearly cost him his job. *Public Safety*, June 1934, **8**, 26–28.
13. Baker, J. S. Accident clinic: Worries caused his accidents. *Public Safety*, July 1934, **8**, 24–26.
14. Baker, J. S. Accident clinic: Could but won't. *Public Safety*, Aug. 1934, **8**, 21 f.
15. Baker, J. S. Accident clinic: Not smart enough for safety. *Public Safety*, Oct. 1934, **8**, 26–28.
16. Baker, J. S. Too long at the wheel. *Public Safety*, Feb. 1935, **9**, 23–26.
17. Baker, J. S. Accident clinic: Discipline would have been futile. *Public Safety*, Sept. 1935, **10**, 24–26.
18. Baker, J. S. Clinical study of accidents. *Public Safety*, June 1936, **11**, 32–34.
19. Baker, J. S. How long on the Highway? *Public Safety*, Jan. 1937, **12**, 34–38.
20. Barton, G. W. Testing automobile drivers—special testing devices. *Proc. 25th Nat. Safety Congr.*, 45–49. Chicago: Nat. Safety Coun., 1937.
21. Berry, D. S. Accident-prone drivers. *Public Safety*, 1937, **12**, 44–46.
22. Billings, C. Science measures the driver's defects. *Nat. Safety News*, July 1934, **30**, 9–11, 48.
23. Bingham, W. V. Personality and

public accidents—a study of accident-prone drivers. *1928 Trans. Nat. Safety Coun.*, 1929, **3**, 174–182.

24. BINGHAM, W. V. The prone-to-accident driver. *Proc. 17th Ann. Conf. Highway Engng.*, Feb. 1931, 23–34.
25. BINGHAM, W. V. Personality and public accidents. *1930 Trans. Nat. Safety Coun.*, 1931, **3**, 140–143.
26. BINGHAM, W. V. Extract from minutes of open meeting of the Committee on the Driver, Nat. Safety Coun., Oct. 13, 1931. *1931 Trans. Nat. Safety Coun.*, 1932, **3**, 24 f.
27. BINGHAM, W. V. The accident prone driver. *Human Factor*, 1932, **6**, 158–169.
28. BINGHAM, W. V. Reliability, validity, and dependability. *J. appl. Psychol.*, 1932, **16**, 116–122.
29. BINGHAM, W. V. Those accident addicts. *Nat. Safety News*, Oct. 1934, **30**, 47, 83.
30. BRAILOVSKII, E. S. Problems in the study of motor reaction in the automobile-driving vocation. *Sovet-Psikhotekh* (*Soviet Psychotechnic*), 1932, **5**, 45–52.
31. BRAILOVSKII, E. S., & SCORODINSKII, G. N. Description of apparatus for studying motor reactions of automobile drivers. *Sovet Psikhotekh* (*Soviet Psychotechnic*), 1932, **5**, 53–58.
32. BRAKEMAN, E. E., & SLOCOMBE, C. S. A review of recent experimental results relevant to the study of individual accident susceptibility. *Psychol. Bull.*, 1929, **26**, 13–38.
33. BRAKEMAN, E. E., & SLOCOMBE, C. S. A readily adaptable apparatus for giving and recording stimuli and responses. *Amer. J. Psychol.*, 1929, **41**, 298–301.
34. BRANSFORD, T. Relation of performance on drivers' tests to automobile accidents and violations of traffic regulations in the District of Columbia. Unpublished doctor's dissertation, American Univ. Washington, D. C., 1939.
35. BURTT, H. E., & FREY, O. C. Suggestions for measuring recklessness, *Person. J.*, 1935, **13**, 39–46.
36. CANTY, A. A note concerning the examination of traffic offenders. *J. appl. Psychol.*, 1936, **20**, 493–498.
37. CANTY, A. What's wrong with violators? Detroit seeks answer with psychoanalysis. *Public Safety*. Mar. 1937, **12**, 15–17.
38. CATTELL, J. McK. Psychological methods to promote highway safety. *Science Monthly*, 1926, **22**, 301–308.
39. COBB, P. W. Selecting the accident prone driver by means of tests. Unpublished report to Highw. Res. Bd., Washington, D. C., Dec. 1938.
40. COBB, P. W. Automobile driver tests administered to 3663 persons in Connecticut, 1936–37, and the relation of the test scores to the accidents sustained. Unpubl. report to Highw. Res. Bd., Washington, D. C., July 1939.
41. COBB, P. W. The limit of usefulness of accident rate as a measure of accident-proneness. *J. appl. Psychol.*, 1940, **24**, 154–159.
42. CRUM, R. W. The highway safety problem. *Proc. 17th ann. meeting Highw. Res. Bd., 1937*, 455–456. Washington: Nat. Res. Coun., 1938.
43. DE SILVA, H. R. *Research on driving skill.* Amherst: Mass. State College, 1935.
44. DE SILVA, H. R. On an investigation of driving skill—1. *Human Factor*, Jan. 1936, **10**, 1–13.
45. DE SILVA, H. R. On an investigation of driving skill—2. *Human Factor*, Feb. 1936, **10**, 50–63.
46. DE SILVA, H. R., & ABERCROMBIE, S. The clinical treatment of traffic violators. *Police J.*, Dec. 1937, **23**, 3–7.
47. DE SILVA, H. R., & FORBES, T. W.

Driver testing results (W.P.A. project 6246–12259.). Cambridge: Harvard Traffic Bur., 1937.

48. De Silva, H. R. Facts about automobile drivers. *Harvard Alumni Bull.*, 1938, **40**, 448–451.
49. De Silva, H. R., & Robinson P. The driver clinic in Delaware high schools. *Safety Educ. Mag.*, Mar. 1938, **17**, 174, 178.
50. De Silva, H. R. Mechanical tests for drivers: Are they of value in promoting safety? *Technology Rev.* (Mass. Institute of Technology) May 1938, **40**, 309–311, 326, 328.
51. De Silva, H. R., & Forbes, T. W. Improving bad drivers. *Safety Engng.*, June 1938, **75**, 13.
52. De Silva, H. R. Age and highway accidents. *Scientific Monthly*, 1938, **47**, 536–545.
53. De Silva, H. R., & Channell, R. Driver clinics in the field. *J. appl. Psychol.*, 1938, **22**, 59–69.
54. De Silva, H. R., & Robinson, P. Light eye and glare sensitivity. *Science*, 1938, **88**, 229.
55. De Silva, H. R., Clafin, R. G. & Simon, W. J. Making safer bus drivers. *Transit J.*, 1938, **82**, 450–451, 471.
56. De Silva, H. R., Frisbee, W. H., & Robinson, P. One-eyed drivers. *Sight-Saving Rev.*, 1938, 8.3, 1–12.
57. De Silva, H. R. Automobile drivers can be improved. *Psychol. Bull.*, 1939, **36**, 284–285.
58. De Silva, H. R. *What we don't know about the driver.* Hartford: Aetna Casualty and Surety Co., circ. 11702.
59. Desrivieres, E., Faillie, R., Jonnard, —, & Vial, —. Reactions psychometrices visuelles en relation avec l'eblouissement par projecteur d'automobiles. *C. r. Acad. sci.*, 1933, **197**, 699–701.
60. Dobrotvorski, N. The principles of psychophysiological tests for aviators. *Vestnik vozdushnogo flota*, 1927, **11**, 32–33.
61. Drake, Charles A. Testing for accident-proneness. Mimeo. copy of paper presented before Amer. Assn. Applied Psychol., Univ. Minn. Aug. 31, 1937.
62. Dyaakova, N. N. Psychotechnical testing of chauffeurs in Moscow. *Psikhofiziol Truda i Psikhotec*, Moscow, 1929.
63. Eddy, R. C. Civil Works Administration in Massachusetts. *Public Safety*, June 1934, **8**, 18 f.
64. Ellis, C. R., Hare, R. A., & Lauer, A. R. The prevalence of visual defects and their relation to automobile driving. *J. Amer. Optometric Assn.* (Offprint circulated by the Association. Vol., date, and original pages not correctly shown.)
65. Emerson, R. W. *The accident-prone employee.* New York: Metropolitan Life Ins. Co., 1929.
66. Farmer, E. The reliability of the criteria used for assessing the value of vocational tests. *Brit. J. Psychol.*, 1933–34, **24**, 109–110.
67. Farmer, E., & Chambers, E. G. A psychological study of individual differences in accident rates. *Industr. Fatig. Res. Bd., Rep. No. 38.* London: H. M. Stationery Office, 1926.
68. Farmer, E., & Chambers, E. G. The prognostic value of some psychological tests. *Industr. Hlth. Res. Bd., Rep. No. 74.* London: H. M. Stationery Office, 1936.
69. Farmer, E., Chambers, E. G. & Kirk, F. J. Tests for accident proneness. *Industr. Hlth. Res. Bd., Rep. No. 68.* London: H. M. Stationery Office, 1933.
70. Farmer, E. Psychological causes of accidents: a critical notice. *Human Factor*, 1937, **11**, 415 ff.
71. Fisher, B. *Mental causes of accidents.*

Boston: Houghton Mifflin Co., 1922.

72. FORBES, T. W. Measuring drivers' reactions. *Person. J.*, 1932, **11**, 111–119.
73. FORBES, T. W. Accidents in traffic and industry as related to the psychology of vision. *Sight-Saving Rev.*, 1936, 6.2, 3–15.
74. FORBES, T. W. Age performance relationships among accident-repeater automobile drivers. *J. consult. Psychol.*, 1938, 2.5, 143–148.
75. FORBES, T. W. A comment on biologically correct traffic equipment, *J. Amer. med. Ass.*, 1938, **112**, 869–870.
76. FORBES, T. W. & MATSON, T. M. Driver judgments in passing on the highway. *J. Psychol.*, 1939, **8**, 3–11.
77. FORBES, T. W. The normal automobile driver as a traffic problem. *J. Gen. Psychol.*, 1939, **20**, 471–474.
78. FORSTER, W. A test for drivers. *Person. J.*, 1928, **7**, 161–171.
79. GEMELLI, A., & PONZO, M. Les Facteurs Psychophysique qui predisposent aux accidents de la rue et les Perspectives d'Organisation Psychotechnique Preventive. *J. de Psychol.*, 1933, **30**, 781–811.
80. GERHARDT, P. W. Scientific selection of employees. *Elec. Railway J.*, 1916, **47**, 943–945.
81. GRADENWITZ, A. Psychology tests for motormen. *Elec. Railway J.* 1922, **59**, 143–146.
82. GREENSHIELDS, B. D. Reaction-time in automobile driving. *J. appl. Psychol.*, 1936, **20**, 353–358.
83. GREENSHIELDS, B. D. Reaction-time and traffic behavior. *Civ. Engng.*, 1937, **7**, 384–386.
84. GREENWOOD, M., & WOODS, H. M. The incidence of industrial accidents upon individuals with special reference to multiple accidents. *Industr. Hlth. Res. Bd., Rep. No. 4.* London: H. M. Stationery Office, 1919.
85. GREENWOOD, M., & YULE, G. U. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or repeated accidents. *J. R. Statis. Soc.*, 1920, **83**, 255–279.
86. HARMAN, N., & BISHOP, M. B. Vision and the motorist. *Practitioner*, 1937, **139**, 218–224.
87. HILDEBRANDT, H. Zur Psychologie die Unfallgefährdeten. *Psychotech. Zeits.*, 1928, **3**, 1–8.
88. HOFFMAN, H. G. Mile-a-minute-men. *Amer. Mag.*, Aug. 1935, **120**, 22–23, 108.
89. HOWARD, R. P. Safety clinics for testing drivers. *Power Wagon.* Nov. 1937.
90. HUSSON, R. Principes de metrologie psychologique (with an introduction by J. M. Lahy). *Actualites scientifiques et industrielles*, No. 555. Paris: Hermann et Cie., 1937.
91. IRWIN, J. O. Correlation methods in psychology. *Brit. J. Psychol.*, 1934, **25**, 86–91.
92. JEKULIN, S. A. Development of first order habits in driving automobiles. Sovet Psikhotekh (Soviet Psychotechnic), 1934, **7**, 138–148.
93. JOHNSON, H. M. The so-called 'coefficient of reliability.' *Proc. 9th Int. Cong. Psychol.*, 1929, 240–241. Princeton: Psychol. Review Co., 1930.
94. JOHNSON, H. M. Born to crash. *Colliers*, July 25, 1936, **98**, 28, 58, 60.
95. JOHNSON, H. M. Tests of "driving skills": What do their exploiters claim they are good for? Unpublished MS (mimeo.), Oct., 1936. In files Highw. Res. Bd., Nat. Res. Counc., Washington, D. C.
96. JOHNSON, H. M. Automobile acci-

dents of youthful drivers. Paper delivered before Section I, A.A.A.S., Indianapolis, 1937. Unpublished MS in files Highw. Res. Bd., Nat. Res. Counc., Washington, D. C.

97. JOHNSON, H. M. The detection of accident-prone drivers. *Proc. 17th ann. meeting, Highw. Res. Bd.*, Dec., 1937, 444–454. Washington: Nat. Res. Counc., 1938.
98. JOHNSON, H. M. The usefulness and limitations of some American tests of drivers' skills. Unpublished MS, read before Highw. Res. Bd., Nat. Res. Counc., 18th ann. meeting, 1938. In file Highw. Res. Bd., Washington, D. C.
99. JOHNSON, H. M. & COBB, P. W. The educational value of "drivers' clinics." *Psychol. Bull.*, 1938, **35**, 758–766.
100. JOHNSON, H. M. Evidence for educational value in drivers' "clinics." *Psychol. Bull.*, 1939, **36**, 674–675.
101. JOHNSON, L. & LAUER, A. R. A study of the effects of induced manual handicaps on automotive performance in relation to reaction time. *J. appl. Psychol.*, 1937, **21**, 85–93.
102. JOHNSON, L. & EVANS, J. E. Apparatus for measuring visual accommodation time to light and to darkness. *J. appl. Psychol.*, 1937, **21**, 705–706.
103. JOHNSON, L. Driver tests and accidents. *Public Safety*, June 1938, **14**, 18–19.
104. JUNG, F. T. Centripetal drift: A fallacy in the evaluation of therapeutic results. *Science*, 1938, **87**, 461–462.
105. KAFKA, G. Zwei neue Apparate zur Eignungsprüfung für Strassenbahne (Vorläufige Mitteilung.) *Z. angew. Psychol.*, 1921, **29**, 95–101.
106. KEELING, S. V. Recent tests for competence in tram driving. *J. Nat'l Inst. Industr. Psychol.*, 1926–27, **3**, 86–93.
107. KEHR, T. Versuchsanordnung zu experimentellen Untersuchung einer kontinuierlichen Aufmerksamkeitsleistung. *Z. angew. Psychol.*, 1916, **11**, 465–479.
108. KING, F. G. W. Tire factors in vehicle control. *Engineering*, (London) Nov. 1, 1935, **140**, 467.
109. LAHY, J. M. *La selection psychophysiologique des travailleurs: conducteurs de tramways et d'autobus.* Paris: Dunod, 1927.
110. LAHY, J. M. La selection psychotechnique des conducteurs de tramways et d'autobus. *Bull. de l'Instit. gen. Psychol.*, **28**, 101–113.
111. LAUER, A. R. Can you pick the safe driver? *Nat. Safety News*, 1931, **24**, 25–27.
112. LAUER, A. R. What types of persons have accidents? *Nat. Safety News*, 1932, **26**, 16 f. 83.
113. LAUER, A. R. How can we measure driving ability? *Nat. Safety News*, 1932, **26**, 64–65.
114. LAUER, A. R. The eyes behind the windshield. *Nat. Safety News*, 1932, **26**, 34–36, 66–67.
115. LAUER, A. R., & KOTVIS, H. L. Automotive manipulation in relation to vision. *J. appl. Psychol.*, 1934, **18**. 422–431.
116. LAUER, A. R. *Manual of tests for automotive operators.* Ames: Iowa State Coll., 1934.
117. LAUER, A. R. Methods of measuring the ability to drive an automobile, Engineering Extension Service, Bull. 115, 35. Ames: Iowa State Coll., 1936.
118. LAUER, A. R. Some practical hints to practitioners on drivers' license examinations. *In 1936 Year Book of Optometry.* Pp. 305–320.
119. LAUER, A. R., & ANDERSON, D. E. An apparatus for measuring changes in bodily resistance. *Amer. J. Psychol.*, 1938, **51**, 156–159.
120. MARBE, K. Praktische Psychologie der Unfälle und Betriebsschäden (Practical psychology of accidents

and industrial injuries). Munch Oldenbourg, 1926.

121. (MARBE, K.) Review of Praktische Psychologie der Unfälle und Betriebsschäden by Karl Marbe. *J. Nat'l Inst. Industr. Psychol.*, 1926–27, **3**, 278–279.
122. MARBE, K. The psychology of accidents. *Human Factor*, Mar. 1935, **9**, 100–104.
123. MARSH, B. W. Fatality hazard much greater for young drivers than for drivers of mature age according to an analysis made by the Safety and Traffic Engineering Dept., A.A.A. Mimeo. copy of report of analysis made by this department, Washington, 1938.
124. MCCANTS, M. Tests used in selecting employees. *Elec. Railway J.*, 1922, **60**, 710–715.
125. MCCANTS, M. Selection and training of employees. *Elec. Railway J.*, 1922, **60**, 679–681.
126. MCCARTER, W. J. A study of accident-proneness of streetcar motormen. Unpublished Master's thesis. Cleveland: Western Reserve Univ., 1932.
127. MILES, G. H. Economy and safety in transport. *J. Nat'l Institute Industr. Psychol.*, 1925, **2**, 192–197.
128. MILES, G. H. The psychology of accidents. *J. Nat'l Inst. Industr. Psychol.*, 1930–31, **5**, 183–192.
129. MILES, G. H., & VINCENT, D. F. The Institute's tests for motor drivers. *Human Factor*, 1934, **8**, 246–257.
130. MILES, G. H. Psychological considerations involved in the application of motor driving tests. *Human Factor*, 1934, **8**, 409–415.
131. MOEDE, W. Zum "Ausbildungkursus in der eignungsprüfung des industrie Lehrlings." *Z. angew. Psychol.*, 1920, **17**, 394–395.
132. MOEDE, W. Psychotechnische Eignungsprüfung in der Industrie. *Prakt. Psychol.*, 1919–1920, **1**, 6–18, 65–81, 339–350, 365–371.
133. MOSS, F. A., & ALLEN, H. H. The personal equation in automobile driving. *J. Soc. Automotive Eng.*, Apr. 1925, **16**, 415–420.
134. MOSS, F. K., & LUCKIESH, M. Nela Park conference on vision. Washington: Safety and Traffic Eng. Dept., A.A.A. 1938. (Mimeographed)
135. MÜNSTERBERG, H. Experiments in the interest of electric railway service, chapter 8, p. 63–82. in *Psychology and industrial efficiency.* New York: Houghton Mifflin, 1913.
136. MUSCIO, B. *Vocational guidance* (A review of the literature). Industr. Hlth. Res. Bd. Report, London: H. M. Stationery Office, 1921.
137. MYERS, C. S. The human factor in accidents. *Human Factor*, 1934, **8**, 266–279.
138. MYERS, C. S. Use of gruesome and humorous propaganda for accident prevention. *Human Factor*, 1936, **10**, 267–272.
139. NEWBOLD, E. M. *A contribution to the study of the human factor in the causation of accidents.* Industr. Hlth. Res. Bd., Rep. No. 34. London: H. M. Stationery Office, 1926.
140. NOVIKOV, V. M. Complex methods of testing fitness for the driving profession. *Soviet Psychotechnic*, 1933, **6**, 333–349.
141. OSBORNE, E. E., VERNON, H. M., & MUSCIO, B. Two contributions to the study of accident causation. *Industr. Hlth. Res. Bd., Rep.* No. 19. London: H. M. Stationery Office, 1922.
142. POTTER, R. D. Tests of reaction time visual acuity, and other physiological factors not sure index of driving fitness, says report to Highway Research Board; use of such tests in granting driving licenses deemed unfair. *Science Service* release, Washington, D. C. Dec. 1, 1938.

143. RAPHAEL, T., LABINE, A. C., FLINN, H. L., & HOFFMAN, L. W. One hundred traffic offenders. *Mental Hygiene*, Oct. 1929, **13**, 809–824.
144. ROLOFF, H. P. Ausbildungskursus in der Eignungsprüfung des industrie Lehrlings, veranstaltet vom Laboratorium für industrielle Psychotechnik in Charlottenburg. *Z. angew. Psychol.*, 1920, **16**, 166–172.
145. SACHS, M. Studien zur Eignungsprüfung der Strassenbahnführer *Z. angew. Psychol.*, 1921, **17**, 199–225.
146. SALMON, T. W. The mind of the operator, *Proc. Yale Univ. Conf. on Vehicle Traffic*, 1924, 50–61.
147. SCHWACKWITZ, A. *Über psychologische Eignungsprüfungen für Verkehrs.* Berlin: Springer, 1920.
148. SCHMITT, E. Unfallaffinität und Psychotechnik im Eisenbahndienst. *Industr. Psychotech.*, 1926, **3**, 144–154, 364–367.
149. SCHORN, M. Unfallaffinität und Psychotechnik. *Industrielle Psychotechnik*, 1924, **1**, 156–160.
150. SELLING, L. S. The physician and the traffic problem. *J. Amer. med. Ass.*, Jan. 9, 1937, **108**, 93–95.
151. SELLING, L. S. The psychological approach to the traffic problem. *Scientific Monthly*, June 1937, **44**, 547–554.
152. SHELLOW, S. M. Selection of Motormen: Further data on value of tests in Milwaukee. *J. Person. Res.*, 1926, **5**, 155–168.
153. SHELLOW, S. M., & MCCARTER, W. J. Who is a good motorman? *Person. J.*, 1927–28, **6**, 338–343.
154. SHUSHAKOV. A. P. Testing the qualifications of locomotive operators in the psychotechnical institute of the Perm railroad by a miniature test. *Zhyrnal psikhologii i isikhotekhniki* (B), 1928, **2**, 14–28.
155. SLOCOMBE, C. S., & BINGHAM, W. V. Men who have accidents. *Person. J.*. 1928, **6**, 251–257.
156. SLOCOMBE, C. S. Unpublished report to National Bureau of Casualty Surety Underwriters, circa 1932.
157. SLOCOMBE, C. S. & BRAKEMAN, E. E. Psychological tests and accident proneness. *Brit. J. Psychol.*, 1930, **21**, 29–38.
158. SLOCOMBE, C. S. It's a habit. *Person. Service Bull.*, 1934, 10.
159. SNOW, A. J. Reduction of automobile accidents by use of psychological tests. *J. Soc. Automotive Eng.*, Aug. 1925, **17**, 163–166.
160. STERN, W. Über eine psychologische Eignungsprüfung für Strassenbahnfahrerinnen. *Z. angew. Psychol.*, 1917, **13**, 91–104.
161. STERN, W. In: Introduction to Sachs, Hildegard. Studien zur Eignungsprüfung der Strassenbahnführer. *Z. angew. Psychol.*, 1920, **17**, 199–225.
162. THONE, F. Machines that measure your skill at the wheel. *Every Week Magazine*, (Sunday Suppl.) Mar. 15, 1936, 3.
163. THONE, F. Youth takes the wheel and the death rate goes up! *Every Week Magazine* (Sunday Suppl.), Feb. 27, 1938, **7**.
164. TOMESCU, P. Examenele Psichotechnice. Romania Medicala (Bucharest) Jan. 1, 1930. In Roumanian. (Abstract by Goldstern, N. *Industr. Psychotech.*, 1930, **7**, 313–315.)
165. TOOPS, H. A., & HAVEN, S. E. Viewing the traffic problem. *J. appl. Psychol.*, 1937, **21**, 185–197.
166. TRAMM, K. A. Die rationelle Ausbildung des Fahrpersonals für Strassenbahnen auf psychotechnischer Brundlage. *Prakt. Psychol.*, 1919–20, **1**, 18–33.
167. TRAMM, K. A. Arbeitswissenschaftliohe Urtersuchüng der menschlich Cerate und Arbeitsverfahren. *Prakt Psychol.*, 1921, **2**, 179–188, 21–219.
168. TRAMM, K. A. Die Psychotechnik im Verkehrwesen. (Editor's abstract of communication to 1921 Internationalen Strassen- und Kleinbahnkongress in Wien.) *Prakt. Psychol.*, 1921, **2**, 354–355.

169. TRAMM, K. A. Psychotechnik und Wirtschaftlichkeit im Strassenbahn. *Prakt. Psychol.*, 1921, **2**, 357–361.

170. TRAMM, K. A. Die Bewahrung des psychotechnischen Prufverfahren für Strassenbahnführer. *Industr. Psychotech.*, 1924, **1**, 36–42.

171. VAN DUZER, W. A. A study of traffic law violations. *Proc. 12th ann. meeting Highw. Res. Bd., Part I.* Washington: Nat. Res. Coun., 1933, 369–378.

172. VAN DUZER, W. A. (Director Dept. Vehicles and Traffic of the District of Columbia) Unpublished charts from reports of WPA project No. 8.

173. VASILEVSKI, S. M. A psychotechnical qualification of autobus chauffeurs. *Psikhofiziol Truda i Psikhotek*, Moscow, 1929.

174. VERNON, H. M. *Accidents and their prevention.* New York: Macmillan Co., 1936. Pp. 44–47, 162–164.

175. VERNON, H. M. Relation of alcohol to road accidents. *Human Factor* 1936, **10**, 255–266.

176. VINCENT, D. F. Institute's apparatus for testing and training drivers. *Human Factor*, 1938, **10**, 64–65.

177. VITELES, M. Research in selection of motormen. I. Survey of the literature. *J. Person. Res.*, 1925, **4**, 100–115.

178. VITELES, M. Research in selection of motormen. II. Methods devised for Milwaukee Electric Railway and Light Co. *J. Person. Res.* 1925, **4**, 173–199.

179. VITELES, M. Psychology in industry. *Psychol. Bull.*, 1926, **23**, 631–680.

180. VITELES, M. Transportation safety by selection and training. *Indus. Psychol.*, 1927, **2**, 119–129.

181. VITELES, M. Psychology in industry. *Psychol. Bull.* 1928, **25**, 309–350.

182. VITELES, M. Psychology in industry. *Psychol. Bull.*, 1930, **27**, 567–635.

183. VITELES, M., & GARDNER, H. M. Weibliche Droschkenchauffeure (Translated by H. Hahn.) *Industr. Psychotech.*, 1931, **8**, 20–26.

184. WECHSLER, D. Tests for taxicab drivers. *J. Person. Res.*, **5**, 24–30.

185. WEISS, A. P., & LAUER, A. R. *Psychological principles in automobile driving.* Columbus: Ohio State Univ., 1931.

186. WHITMER, C. C. The relationship between personality and accidents, *1929 Trans. Nat'l Safety Coun.*, 1930, **3**, 97–101.

187. WILLIAMS, S. J. Teaching old drivers new tricks. *Public Safety*, May 1937, **12**, 24–25, 58.

188. WILSON, O. W. Wichita traffic clinic. *Public Safety*, Aug. **13**, 8–9, 46.

189. YERKES, R. M. (Ed.) Psychological examining in the United States Army. *Mem. Nat. Acad. Sci.*, 1921, 15.

190. YORDAN, E. L. The accident jinx can be broken. *Public Safety*, Apr. 1937, **13**, 26–27.

191. YULE, G. U. The function of statistical method in scientific investigation. *Industr. Hlth. Res. Bd., Rep.* No. 28.

192. YULE, G. U., & KENDALL, M. G. *An introduction to the theory of statistics.* London: Griffin. 1927.

193. *The accident-prone employee—a study of electric railway operation.* Cleveland Railway Co. (with the cooperation of the Policy Holder Bur. of the Metropolitan Life. Ins. Co.).

194. *AAA driver rating manual.* Amer. Auto. Ass., undated booklet.

195. *AAA driver testing equipment on nation-wide tour.* Washington: Amer. Auto. Assn., undated booklet, circulated 1936.

196. *Analysis of 16,500 road accidents in Connecticut.* (Anon.) *Engng. News-Rec.*, 1924, **92**, 807.

197. *Annual statistical report 1937.* California (State of) Dept. of Motor Vehicles. Sacramento: Bur. of Statistics, undated booklet, circulated Apr. 1937.

198. *Anti-accident diet.* (Anon.) *Public Safety*, Apr. 1937, **12**, 56.

199. *Are you a safe driver?* Iowa State

Motor Vehicle Dept. Undated booklet, circulated 1936.

200. *The current program of the Bureau for Street Traffic Research.* Bur. for Street Traffic Res. Cambridge: Harvard Univ., 1936.
201. *Death on the highways.* (Anon.) *Look*, Dec. 6, 1938, **2**, 5–11.
202. *Drinking drivers tests.* (Anon.) *Public Safety*, May 1936, **10**, 24.
203. *Examining applicants for drivers' licenses.* Chicago: Nat. Safety Counc., Inc. 1934.
204. *Good vision essential for vehicle operators.* (Anon.) *Public Safety*, Nov. 1932, **11**, 24 f.
205. *Highway safety demonstration.* Hartford: Aetna Casualty and Surety Co. Undated booklet, circulated 1936.
206. *How good a driver are you?* Hartford: Aetna Casualty and Surety Co. 1936.
207. *Job analysis of highway safety.* Personnel Research Federation. *Person. J.*, May 1938, **17**, 25–30.
208. *Keystone safety test trailer.* Keystone Auto Club. Undated folder.
209. *Let's be skillful.* Hartford: Aetna Casualty and Surety Co. Undated booklet, circulated 1936.
210. *Making the highways safe.* (Anon.) *J. Soc. automot. Engrs.*, July 1925, **17**, 13.
211. *Motor-vehicle traffic conditions in the United States.* Part 6: *The accident-prone driver.* U. S. Secretary of Agriculture. (House Doc. 462, 75th Congress, 3rd Session.) Washington: U. S. Govt. Printing Office, 1938.
212. *Note sur la selection psychotechnique du personnel à la S.T.C..PP.* S.T.C.R.P. Paris: G. Fuseau, 1937. Distributed by Societe des transports en commun de la region parisienne.
213. *Novel device registers reaction time.* (Anon.) *Public Safety*, June 1935, **9**, 14 f.
214. *Odd driving tests seek the causes of auto accidents.* (Anon.) *Pop. Sci. Mon.*, 1940, **137**, 135 f.
215. *Preventing Taxicab accidents.* New York: Metropolitan Life Ins. Co., undated, circulated 1931.
216. *Preliminary report of Traffic Safety Committee Council paper 105 of 1936.* Trinidad and Tobago: A. L. Rhodes, Government Printer, 1936.
217. *Reaction time.* (Anon.) *Public Safety*, January 1936, **10**, 26 f.
218. *Report on fatal road accidents which occurred during the year 1933.* Ministry of Transport (Great Britain). London: H. M. Stationery Office, 1936.
219. *Report on fatal road accidents which occurred during the year* **1935**. Ministry of Transport (Great Britain). London: H. M. Stationery Office. 1936.
220. *Report of a state-wide survey of operators' tests and comprehensive study of 1935 fatalities with special attention to records of "repeaters" as compiled by WPA project 1603.* Hartford: Connecticut Dept. of Motor Vehicles. 1937.
221. *Safe transportation.* New York: Personnel Research Federation, undated pamphlet, circulated 1930.
222. *Selection tests on the German railways* (based on lecture by Dr. Glasel Dresden, May 1926). (Anon.) *J. Nat'l. Inst. Industr. Psychol.*, 1926–1927, **3**, 201–204.
223. *La service moderne d'embauchage de la S.T.C.R.P.* ST.C.R.P. Paris: Maulde et Renou. 1937.
224. *Sate of Illinois Safety bulletin.* Illinois. April 1, 1936, 2.
225. *Will the slaughter go on?* Personnel Research Federation. *Person. J.*, Apr. 1938, **16**, 333–339.
226. *Your rating as a safe driver as tested at the safe driver clinic of the Chicago Motor Club.* Chicago Motor Club. Undated circular.

A GENERAL TEST FOR TREND*

HOWARD W. ALEXANDER
Adrian College

In psychological or biological research an experimental design is frequently used in which a number of individuals (humans or animals) are subjected to a series of trials, all of the individuals being assumed to be tested simultaneously. In many psychological experiments one variable is measured while another takes a series of specified values, as in experiments on learning or conditioning, on work decrements, on dark adaptation and the effects of different degrees of illumination, or on visual span of apprehension and report. The statistical treatment of such data has generally been confined to a comparison of the experimental groups either at a specified point of the experiment, or with respect to increment or decrement over a specified period. Thus only one or two points of the curves are actually used in the analysis, and often the great bulk of the data remains unused. The present paper will deal with methods enabling the experimenter to employ the data of his complete series of trials.

The complete set of data will usually take the form of a table of individuals by trials. For example, we may wish to determine the reliability of a mental test, in which case the trials are presumed to be under constant conditions and the ability of the individuals is presumed not to have changed in the interval between the trials. Again, the trials may be spread through a period of experimental stress (dietary, environmental, psychological) and the object may be to determine whether the individuals show a consistent and measurable reaction to the stress. Or the subjects may be divided into several groups, each maintained on its own experimental regime, and we may wish to decide whether the groups exhibit different patterns of response.

Often we have little *a priori* basis for expecting any particular pat-

* From the Laboratory of Physiological Hygiene, University of Minnesota. This work was supported in part under the terms of a contract between the Regents of the University of Minnesota, and the Office of Scientific Research and Development. Important financial assistance was also provided by the Nutrition Foundation, Inc., the U. S. Sugar Cane Refiners' Association, N. Y., the Corn Industries Research Foundation, N. Y., Swift and Co., Chicago, the National Confectioners' Association, the National Dairy Council, Chicago, and the Graduate Medical Research Fund, University of Minnesota. The author wishes to express his indebtedness to all the members of the staff of the Laboratory of Physiological Hygiene, and especially to Mr. Harold Guetzkow and Mr. Richard Seymour.

tern of response to the imposed stress. The problem then is to distinguish genuine response from random fluctuation. One criterion that will be used in this paper is consistency of performance from individual to individual, which is exhibited graphically when the performance graphs tend to be parallel. The general method will allow the rate of response to differ from individual to individual, while including a component common to the group.

In the case of two trials, we may test whether there has been a consistent change in performance from the first to the second trial by using Student's test for paired variates (**6**, pp. **43–45**). This case has been thoroughly dealt with by R. W. B. Jackson in connection with his development of measures of reliability (**3**, **5**). He uses the term "trial effect" to denote a significant difference of the means of the two trials, and he supplies a test for trial effect (equivalent to Student's test) to be made prior to the estimation of reliability. In the Appendix of his monograph on reliability (**5**) he shows that the maximum likelihood estimate of reliability is dependent upon the presence or absence of trial effect.

In the present paper we shall use the term *trend* to denote significant trial-to-trial fluctuation, and we shall develop a test for trend which is a generalization of the simple test supplied by Student's test (or Jackson's equivalent method) in the case of two trials.

Basic Definitions and Assumptions

Consider a table consisting of the scores of n individuals on k trials of a certain test, the trials being spread over an experimental period of considerable length. Our problem is to determine whether the trial means deviate significantly from the general mean. If the trial means fluctuate about the general mean to a greater degree than can be accounted for by random variation, we shall say that *trend* is present; we wish to devise tests of significance for trend.

We assume that the trials could have been taken so close together that values are available for every moment of the experimental period. Furthermore, we assume that the experiment is indefinitely repeatable with the same individuals, and that the individuals do not change between repetitions. Thus, for each moment of the experiment, and for each individual, we obtain a distribution of repetitions. The mean value of this distribution may be graphed against time for each individual, and we thus obtain an *individual regression curve*. On the other hand, the n individuals of the group constitute a sample from a population of like individuals, so that we may conceive of a population of individual re-

gression curves, and a population mean of such curves, the *group regression curve.*

Consider again the distribution of repetitions for a single individual at a single moment of the experiment. We shall assume that this distribution is always normal, and that it has a variance σ^2 which is the same for all individuals throughout the experiment.* We further assume

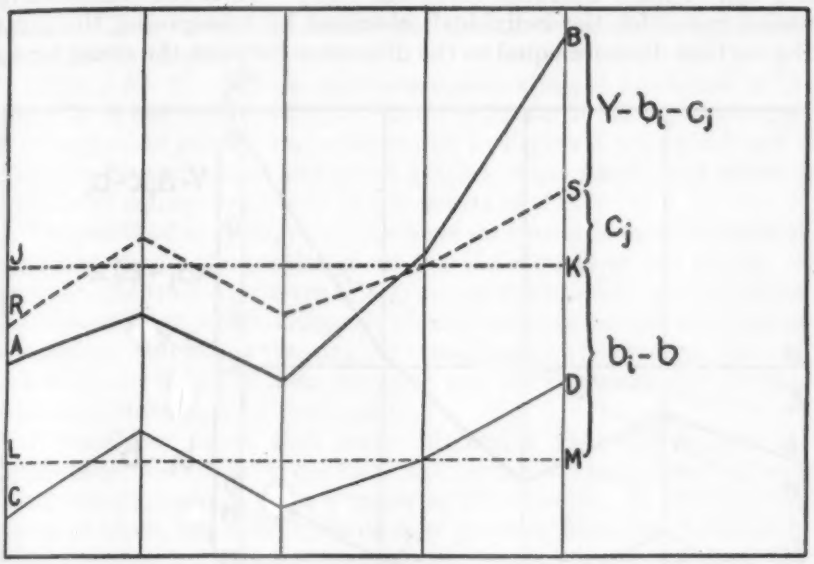

FIGURE 1. ESTIMATION METHOD A.

that any pair of such distributions are independent or uncorrelated. The procedure in testing for trend will be to obtain an estimate s^2 of the random or error variance σ^2 and use it, in the form of an F-test, to evaluate other components of variation. We shall, in effect, be testing whether the group regression curve is other than a horizontal straight line.

METHODS OF ESTIMATION

In order to obtain an estimate of random or error variance, we require to have a method of estimating certain points on the individual regression curves, so that the deviations of the given values from the

* This is known as the assumption of homogeneity of variance, and it should be tested when there is doubt as to its validity. Tests for this purpose have been provided by Bartlett (see Snedecor 6, pp. 249–251) and by Welch (see Jackson, 4, pp. 40–41).

estimates may be obtained. We have several devices at our disposal, which we may use alone or in combination.

Method A. We may use the trial means, which estimate the group regression curve, as a basis for estimating the individual regression curves. Figure 1 illustrates the case of five trials ($k=5$), in which AB is the graph of the 5 trial values for a particular individual, and CD is the graph of the 5 trial means for the group. RS is the estimated regression curve for the individual, obtained by transposing the graph CD a vertical distance equal to the difference between the mean for the

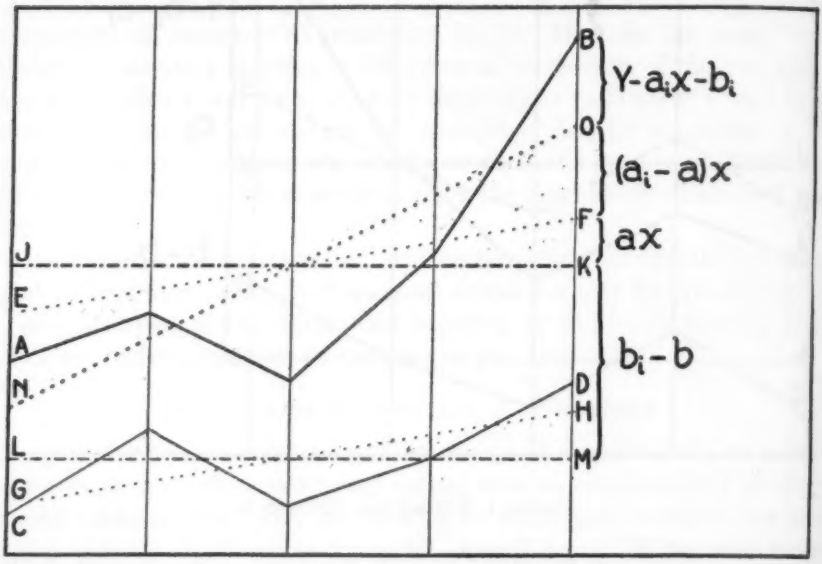

FIGURE 2. ESTIMATION METHOD B.

individual and the general mean. This method of estimation may be expressed in another way, which will be useful later on. LM is a horizontal line whose ordinate is equal to the general mean. The trial fluctuation is measured by the deviations of the 5 trial means from the line LM. The line JK is likewise horizontal, and with ordinate equal to the mean for the individual whose graph is AB. We may say that RS is constructed so that its deviations from JK are equal to the corresponding deviations of CD from LM.

It is clear that under this method of estimation, all the estimated curves for the n individuals will be *parallel*, and the suitability of this method will depend on whether the individual curves can well be represented as parallel lines. If there is reason to suspect that the curves

deviate significantly from parallelism, then this method should not be used; such might be the case, for example, when a number of individuals are going through a period of training with respect to the test which is being analysed, since we know that training takes place at different rates for different individuals. On the other hand, an example will be given later in which a group of individuals under a controlled laboratory situation show a trend which is common to all the individuals of the group, with no significant differences in trend for the different individuals. Under such circumstances, the above method of estimation would be justified.

Method B. We may use the least square straight line fitted to the k points for a particular individual as an estimate of his regression curve. This method of estimation is illustrated in Figure 2, where AB and CD are again the individual and group graphs, respectively, and where NO is the least square line fitted to the points of AB.

This method is applicable if we have no reason to suspect consistent curvilinearity in the graphs of all the individuals of the group. For example, the training curves of a group of individuals may be taken as approximately straight (although with different slopes) at the beginning of training. But as the training curves approach the plateau, there may be a tendency for all of them to curve, and the representation by means of straight lines may be inadequate.

It should be noted that under Method A the time scale is of no significance; we use only the fact that certain measures have been obtained simultaneously on a number of individuals. In Method B, on the other hand, the time variable is of great significance. Ordinarily it would be measured in days, weeks or other units of calendar time. However, there may be instances in which it is more sensible, from the standpoint of psychology, to measure time in terms of the number of trials, ignoring the calendar time.

Method C. We may use a method which combines features of the two preceding ones, permitting the individual estimates to have different slopes, and at the same time allowing for a curvilinear trend common to all individuals. This method of estimating is illustrated in Figure 3. AB and CD are again the individual and group graphs, while NO and GH are least square straight lines fitted to the points of the graphs. We construct the estimated graph PQ in such a way that the deviations of PQ from NO are equal to the corresponding deviations of CD from GH. These deviations thus constitute the portion of the estimate that is common to all the individuals, and which may represent, for example, some fluctuation in the experimental environment, or a common curvature on approaching the practice plateau.

Other methods of estimation could be employed, such as those involving polynomials or other mathematical curves. Such estimations go beyond the scope of the present paper.

Analytical Formulation

Let x designate the time variable, represented as the abscissa in the Figures, and let Y designate the experimental values, represented as ordinates. The formulas are greatly simplified if the origin of x is so chosen that $\Sigma x = 0$, where the summation is over the k trials. In keeping with statistical convention, we shall use small letters for x (since the

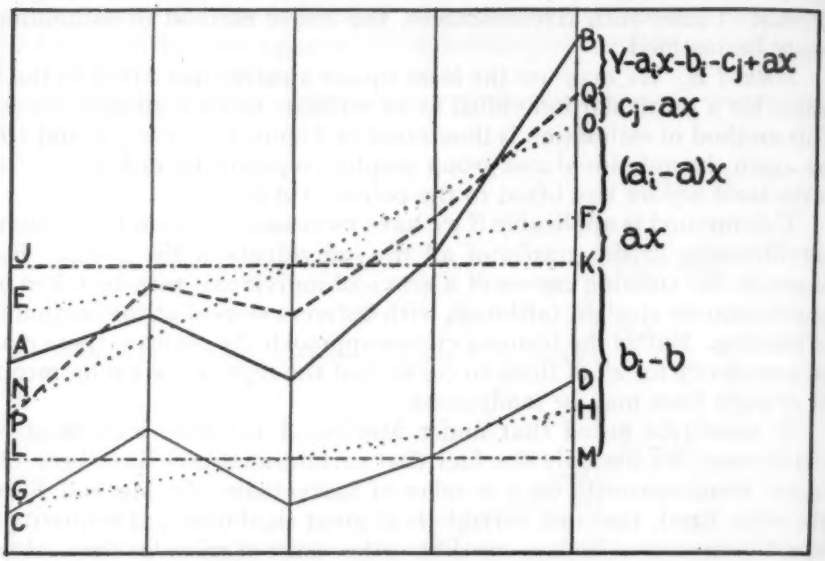

Figure 3. Estimation Method C.

origin is the mean value) and capitals for Y. For the purposes of the tests of significance the time units are immaterial. Let

P_i (or simply P) $= \Sigma xY$
I_i (or simply I) $= \Sigma Y$ } summed for the ith individual

T_j (or simply T) $= \Sigma Y$, summed for the jth trial,

GT = Grand total of the table

$c_j = T_j/n - GT/kn$ = Deviation of trial mean from general mean.

Let $Y = a_i x + b_i$ designate the least square line for a single individual, and let $Y = ax + b$ be the least square line for the whole group. Then a_i, b_i, a and b are given by

$$a_i = P_i/\Sigma x^2, \qquad b_i = I_i/k, \qquad a = \Sigma a_i/n, \qquad b = \Sigma b_i/n.$$

Thus b_i is simply the mean of the values for a single individual, while b is the general mean.

Following the methods of the analysis of variance, we shall separate the total sum of squares, $\Sigma(Y-b)^2$, into a number of components, each corresponding to a separate source of variation. In Figures 1, 2 and 3, the essential breakdown of the analysis is shown on the right hand side of the figure, each component being indicated by a bracket.

In Figure 1, the deviations of the trial means from the general mean are represented by the deviations of CD from the horizontal line LM;) analytically they are represented by the quantities c_j defined above. Since the line JK has constant ordinate b_i ,the estimated points of the graph RS have ordinate b_i+c_j. Hence

$$\begin{aligned} Y - b &= BM = BS + SK + KM \\ &= (Y - b_i - c_j) + c_j + (b_i - b). \end{aligned}$$

Thus

$$\begin{aligned} (Y - b)^2 = (Y - b_i - c_j)^2 + c_j^2 + (b_i - b)^2 + 2(Y - b_i - c_j)c_j \\ + 2(Y - b_i - c_j)(b_i - b) + 2c_j(b_i - b). \end{aligned}$$

When all terms of this equation are summed over all values of Y, it can be shown that the last three terms (the cross-product terms) vanish. This will happen also with the other methods of estimation. Thus

$$\Sigma(Y - b)^2 = \Sigma(Y - b_i - c_j)^2 + \Sigma c_j^2 + \Sigma(b_i - b)^2.$$

This partitioning of the total sum of squares is identical with that obtained in the usual analysis of variance of a two-way table (6, Chap 11), and the three terms listed above are, respectively, the sums of squares associated with "interaction," with "between trials" variation and with "between individuals" variation. The interaction term is the estimate s^2 of the error variance afforded by this analysis. It might have been called "deviations from estimation," but for clarity this term will be reserved for the error term under Method C while interaction will be used for the present error term, in order that the various estimates of error may later be compared. We shall have to deal with "individual slopes" later, so we shall use the phrase "between individual means" to designate the last component, rather than "between individuals." In the table below, these three terms are listed, and suitable computational expressions are given for the sums of squares.

Source of variation	*Sum of squares*	*Degrees of freedom*	*Mean square*
Interaction	$\Sigma Y^2-\Sigma I^2/k-\Sigma T^2/n+GT^2/nk$	$(n-1)(k-1)$	V_{ii}
Beween trials	$\Sigma T^2/n-GT^2/nk$	$k-1$	V_t
Bet. individual means	$\Sigma I^2/k-GT^2/nk$	$n-1$	V_{im}
Total	$\Sigma Y^2-GT^2/nk$	$nk-1$	

To test for trend under this method of analysis, we determine whether V_t is significantly greater than V_{it} by means of the ratio $F = V_t/V_{it}$. If this ratio is significant, we conclude that in the trial-to-trial fluctuation there is a significant component common to all the individuals of the group, to which we have given the name *trend*.

Method B: In Figure 2, NO has the equation $Y = a_i x + b_i$, while GH has the equation $Y = ax + b$. Let EF be a line through the intersection of ON and JK, and parallel to GH; its equation will be $Y = ax + b_i$. The lines JK and LM, with constant ordinates b_i and b, respectively, will have equations $Y = b_i$ and $Y = b$. Thus

$$\begin{aligned} Y - b = BM &= BO + OF + FK + KM \\ &= (Y - a_i x - b_i) + (a_i - a)x + ax + (b_i - b) \end{aligned}$$

and

$$\Sigma(Y - b)^2 = \Sigma(Y - a_i x - b_i)^2 + \Sigma x^2 \, \Sigma \, (a_i - a)^2 + a^2 \Sigma x^2 + \Sigma(b_i - b)^2.$$

These terms may be computed, and the corresponding mean squares obtained, as indicated in the table below:

Source of variation	*Sum of squares*	*Degrees of freedom*	*Mean square*
Individual deviations from linearity	$\Sigma Y^2 - \Sigma I^2/k - \Sigma P^2/\Sigma x^2$	$n(k-2)$	V_{idfl}
Between individual slopes	$\Sigma P^2/\Sigma x^2 - (\Sigma P)^2/n\Sigma x^2$	$n-1$	V_{is}
Group slope	$(\Sigma P)^2/n\Sigma x^2$	1	V_{gs}
Between individual means	$\Sigma I^2/k - GT^2/nk$	$n-1$	V_{im}
Total	$\Sigma Y^2 - GT^2/nk$	$nk-1$	

The basic test for trend in this type of analysis is provided by the ratio $F = V_{gs}/V_{idfl}$, which constitutes a test of the hypothesis that the group slope is zero. Although a complete discussion of the various possibilities of even so simple an analysis is not possible in the present account, it is worth while to point out that the term V_{is} really has considerable bearing upon the above test. If V_{gs}/V_{idfl} and V_{gs}/V_{is} are both significant, no difficulty arises, since there is evidently a component of group slope which is significantly different from zero, in addition to sources of variation present in either V_{is} or V_{idfl}. On the other hand, if V_{gs}/V_{idfl} is significant while V_{gs}/V_{is} is non-significant, the implication is that although the group slope term is not attributable to random error, it may be attributable to the variation of the slope from individual to individual, and one would conclude that there is no significant slope common to all the individuals. Instances will not infrequently

arise in which there are ambiguities in the tests of significance which cannot be resolved by methods at present available.

Method C: Under this method of estimation, we use the deviations of the group graph, CD, from the corresponding least square line, GH, in order to adjust the estimates obtained by fitting least square lines to the data for the individuals. The deviations of CD from GH are given by $c_j - ax$; these values are to be added to the linear estimates given by $Y = a_i x + b_i$, giving rise to the estimated graph PQ (Figure 3) whose equation is $Y = a_i x + b_i + c_j - ax$. The other lines in Figure 3 have the same meanings as in Figures 1 and 2. We now separate $Y - b$ into five components:

$$\begin{aligned} Y - b = BM &= BQ + QO + OF + FK + KM \\ &= (Y - a_i x - b_i - c_j + ax) + (c_j - ax) \\ &\quad + (a_i - a)x + ax + (b_i - b). \end{aligned}$$

and

$$\begin{aligned} \Sigma(Y - b)^2 = {} & \Sigma(Y - a_i x - b_i - c_j + ax)^2 + \Sigma(c_j - ax)^2 \\ & + \Sigma(a_i - a)^2 \Sigma x^2 + \Sigma a^2 x^2 + \Sigma(b_i - b)^2. \end{aligned}$$

The computations for the analysis are indicated below:

Source of variation	*Sum of squares*	*Degrees of freedom*	*Mean square*
Individual deviations from estimation	$\Sigma Y^2 - \Sigma T^2/n - \Sigma I^2/k + GT^2/nk - \Sigma P^2/\Sigma x^2 + (\Sigma P)^2/n\Sigma x^2$	$(n-1)(k-2)$	V_{idfe}
Group deviations from linearity	$\Sigma T^2/n - GT^2/nk - (\Sigma P)^2/n\Sigma x^2$	$k-2$	V_{gdfl}
Between individual slopes	$\Sigma P^2/\Sigma x^2 - (\Sigma P)^2/n\Sigma x^2$	$n-1$	V_{is}
Group slope	$(\Sigma P)^2/n\Sigma x^2$	1	V_{gs}
Between individual means	$\Sigma I^2/k - GT^2/nk$	$n-1$	V_{im}

Here the estimate of the random or error variation, s^2, is given by V_{idfe}, and we have two distinct tests for trend

$F = V_{gs}/V_{idfe}$	Tests whether the group as a whole is tending to increase or decrease in linear fashion
$F = V_{gdfl}/V_{idfe}$	Tests whether the group is tending to fluctuate about the linear trend line in a consistent fashion.

Ambiguous cases can arise with the present method, similar to those discussed under Method B, but it is not possible to deal with them completely in the present account.

It is useful to note that all the terms in the two earlier tables can be obtained by re-combining components from the above table. Thus the following equations can be set up, applying to both sums of squares and degrees of freedom:

$$\left\{\begin{matrix}\text{Individual deviations}\\ \text{from linearity}\end{matrix}\right\} = \left\{\begin{matrix}\text{Individual deviations}\\ \text{from estimation}\end{matrix}\right\} + \left\{\begin{matrix}\text{Group deviations}\\ \text{from linearity}\end{matrix}\right\}$$

$$\text{(Interaction)} = \left\{\begin{matrix}\text{Individual deviations}\\ \text{from estimation}\end{matrix}\right\} + \left\{\begin{matrix}\text{Between individual}\\ \text{slopes}\end{matrix}\right\}$$

$$\text{(Between trials)} = \left\{\begin{matrix}\text{Group deviations}\\ \text{from linearity}\end{matrix}\right\} + \text{(Group slope)}$$

Notice that the term "individual deviations from estimation," which is the estimate of error variance under Method *C*, appears as a component in each of the earlier estimates of error variance, namely, "interaction" under Method A and "individual deviations from linearity" under Method B.

Numerical Example

In order to illustrate and further clarify the concepts of the preceding sections a fictitious numerical example will be presented. It was

TABLE 1

		Trials 1	2	3	4	5	*Total* $=I$	Mean	P ΣxY
Individuals	1	300	190	200	330	280	1300	260	100
	2	320	320	380	360	270	1650	330	−60
	3	280	350	230	290	400	1550	310	180
	4	300	300	230	380	290	1500	300	60
Group 1	$\Sigma T_1=$						2950		$\Sigma_1 P=$ 40
	Total $=T_1$	620	510	580	690	550		295	20
	Mean	310	255	290	345	275	$\Sigma T_2=$		$\Sigma_2 P=$
Group 2	*Total* $=T_2$	580	650	460	670	690	3050		240
	Mea	290	325	230	335	345		305	120
Overall	$GT=$						6000		$\Sigma P=$ 280
	Total $=T$	1200	1160	1040	1360	1240		300	70
	Mean	300	290	260	340	310			

constructed from two-digit random numbers, which were somewhat adjusted and finally multiplied by 10 to make for computational convenience. The table will be regarded as representing the scores of four individuals on five trials. In order to illustrate the methods for the comparison of the trends of two groups, which will be developed in later

sections, the four individuals are divided into two groups with two individuals in each.

A suitable time scale is required for the construction of the final column of $P = \Sigma xY$. Assuming the five trials to be equally spaced with respect to time, we take the time interval between trials to be unity. We also wish to have $\Sigma x = O$, since the formulas are made simpler by this fact. These considerations lead us to the values -2, -1, 0, 1 and 2 for x, and we find $\Sigma x^2 = 10$. The column of P is simply the sum of the products xY computed for each row.

The sums of squares required for the analyses under Methods A, B and C are as follows:

$$\Sigma Y^2 = 1{,}864{,}400 \qquad \Sigma I^2/5 = 1{,}813{,}000 \qquad \Sigma P^2/10 = 4{,}960$$
$$\Sigma T^2/4 = 1{,}813{,}600 \qquad GT^2/20 = 1{,}800{,}000 \qquad (\Sigma P)^2/40 = 1{,}960$$

When these values are substituted in the analysis tables for Methods A, B and C we obtain the analyses given in Tables 2, 3 and 4.

TABLE 2: METHOD A

Source of variation	*Sum of squares*	*df*	*Mean square*
Interaction	37,800	12	$V_{it} = 3{,}150$
Between trials	13,600	4	$V_t = 3{,}400$
Between individual means	13,000	3	$V_{im} = 4{,}333$
Total	64,400	19	

TABLE 3: METHOD B

Source of variation	*Sum of squares*	*df*	*Mean square*
Individual deviations from linearity	46,440	12	$V_{id/l} = 3{,}870$
Between individual slopes	3,000	3	$V_{i_s} = 1{,}000$
Group slope	1,960	1	$V_{g_s} = 1{,}960$
Between individual means	13,000	3	$V_{im} = 4{,}333$
Total	64,400	19	

TABLE 4: METHOD C

Source of variation	*Sum of squares*	*df*	*Mean square*
Individual deviations from estimation	34,800	9	$V_{id/e} = 3{,}867$
Group deviations from linearity	11,640	3	$V_{gd/l} = 3{,}880$
Between individual slopes	3,000	3	$V_{i_s} = 1{,}000$
Group slope	1,960	1	$V_{g_s} = 1{,}960$
Between individual means	13,000	3	$V_{im} = 4{,}333$
Total	64,400	19	

The tests for trend are as follows:

Method A: $F = V_t/V_{it} = 3{,}400/3{,}150 = 1.08$

Method B: $F = V_{os}/V_{idfl} = 1{,}960/3{,}870 = .51$

Method C: $F = V_{odfl}/V_{idfe} = 3{,}880/3{,}867 = 1.00$

$F = V_{os}/V_{idfe} = 1{,}960/3{,}867 = .51$

None of these ratios approaches significance, and the conclusion is that no trend is present in the table. Inasmuch as the table was constructed from random numbers, this conclusion is to be expected. In the language of the analysis of variance, one may say that the null hypothesis has been sustained that only one source of variation is present in the original table.

Computation of Estimated Values

It is not difficult to exhibit tables of estimated values corresponding to each method of estimation. This is not a practical procedure for empirical examples, since all that is normally required for an analysis is the table of analysis of variance, computed as indicated above, together with some examination of the means of the rows and columns. However, the underlying ideas may be clarified by such tables of estimates, especially when we come to deal with the somewhat more complicated case of the comparison of group trends.

It is first necessary to indicate how *linear estimates* corresponding to any row of Table 1 may be obtained. The slope of the least square straight line fitted to the five trial values plotted against time (as in Fig. 2) is given by $\Sigma xY/\Sigma x^2 = P/10$. This line intersects the Y axis at a point whose ordinate is $\Sigma Y/k = T/5$. Thus its equation is

$$Y = (P/10)x + T/5.$$

For the first row of Table 1 we have a slope of $P/10 = 100/10 = 10$, and a mean value of $T/5 = 260$. Using these values together with the time values $x = -2, -1, 0, 1, 2$, we obtain the linear estimates 240, 250, 260, 270, 280.

All the tables required for the construction of the estimates in which we are interested are contained in the comprehensive Table 5, which includes the 9 sub-tables $A, B, C, \ldots, J$. Table A is identical with the original part of Table 1. Table B consists of the linear estimates of each row of Table A. Table C consists of the mean values of the rows repeated 5 times. Tables D, E and F are obtained from those immediately above them by averaging the first two rows of the upper table and writ-

ing the five values thus obtained in the first and second rows of the lower table; and similarly with the last two rows. Finally, Tables G, H and J are obtained from Tables A, B and C by averaging all four items of each column of the upper tables, and repeating these averages four times in the corresponding column of the lower table. Many of these values

TABLE 5

	Trial values	*Linear estimates*	*Means*
Individuals	Table A Ind. trial values 300 190 200 330 280 320 320 380 360 270 280 350 230 290 400 300 300 230 380 290 $ss = 1{,}864{,}400$ $df = 20$	Table B Ind. linear est. 240 250 260 270 280 342 336 330 324 318 274 292 310 328 346 288 294 300 306 312 $ss = 1{,}817{,}960$ $df = 8$	Table C Individual means 260 260 260 260 260 330 330 330 330 330 310 310 310 310 310 300 300 300 300 300 $ss = 1{,}813{,}000$ $df = 4$
Groups	Table D Group trial values 310 255 290 345 275 310 255 290 345 275 290 325 230 335 345 290 325 230 335 345 $ss = 1{,}827{,}500$ $df = 10$	Table E Group linear est. 291 293 295 297 299 291 293 295 297 299 281 293 305 317 329 281 293 305 317 329 $ss = 1{,}803{,}460$ $df = 4$	Table F Group mean 295 295 295 295 295 295 295 295 295 295 305 305 305 305 305 305 305 305 305 305 $ss = 1{,}800{,}500$ $df = 2$
Overall	Table G Overall trial values 300 290 260 340 310 300 290 260 340 310 300 290 260 340 310 300 290 260 340 310 $ss = 1{,}813{,}600$ $df = 5$	Table H Overall linear est. 286 293 300 307 314 286 293 300 307 314 286 293 300 307 314 286 293 300 307 314 $ss = 1{,}801{,}960$ $df = 2$	Table J Overall mean 300 300 300 300 300 300 300 300 300 300 300 300 300 300 300 300 300 300 300 300 $ss = 1{,}800{,}000$ $df = 1$

have already been exhibited in Table 1. Note that Table E gives the linear estimates corresponding to Table D; a similar relation holds between Tables G and H.

Each of the 8 tables B, C, . . . , J is a type of estimate of Table A. Other types of estimates may be formed by a process of combining these 8 tables. The method of combination may be illustrated by means of

Table (C+G−J), which contains the estimates of Table A according to estimation Method A. It is formed, as its title indicates, by adding together corresponding elements of Tables C and G and subtracting the corresponding element of Table J. For example, take the upper right hand elements of these three tables. Table C. 260, Table G. 310, Table J. 300. The estimate obtained from these is 260+310−300=270, which appears in the upper right of Table $(C+G-J)$. It was remarked earlier that with Method A of estimation the individual estimated graphs would be parallel. This can easily be verified for the values of Table $(C+G-J)$.

TABLE $(C+G-J)$

260	250	220	300	270
330	320	290	370	340
310	300	270	350	320
300	290	260	340	310

Below each of the sub-tables of Table 5 is given the sum of squares (*ss*) and the degrees of freedom (*df*). The sum of squares is simply the sum of the squares of the 20 items in the corresponding table. The degrees of freedom may be defined as the number of independent numbers in the table, or the number of numbers required to specify the table. Thus the 20 items of Table A are independent, and the *df* is 20. Each row of Table B is specified by its mean value and its slope, so the *df* is 8. Similar considerations lead to the other *df* values.

The following useful rule may be stated. In any table formed by combining the sub-tables of Table 5, *the sum of squares and the degrees of freedom follow the same law of combination as the individual items.* For example, the sum of squares of Table $(C+G-J)$ is found, by direct computation, to be 1,826,600, which we may designate by $ss(C+G-J)$. For the three sub-tables we have $ss(C)=1{,}813{,}000$, $ss(G)=1{,}813{,}600$, $ss(J)=1{,}800{,}000$. From these values it is easy to verify that

$$ss(C+G-J)=ss(C)+ss(G)-ss(J).$$

Similarly, to obtain the *df* of Table $(C+G-J)$, which we may denote by $df(C+G-J)$, we use

$$df(C+G-J)=df(C)+df(G)-df(J)=4+5-1=8.$$

It can be verified that exactly 8 numbers are required to specify completely the Table $(C+G-J)$.

The method of combining tables may be used to exhibit a table of components corresponding to each source of variation entering into

Methods A, B and C. The actual tables need not be given here; the formulas for obtaining them are given in Table 6. It is easy to verify that the formulas yield the values for the sums of squares and *df* already listed in Tables 2, 3 and 4.

TABLE 6

Source of variation	*Formula*
Interaction	$(A-G-C+J)$
Between trials	$(G-J)$
Between individual means	$(C-J)$
Individual deviations from linearity	$(A-B)$
Between individual slopes	$(B-C-H+J)$
Group slope	$(H-J)$
Individual deviations from estimation	$(A-B-G+H)$
Group deviations from linearity	$(G-H)$

Each of the formulas given in Table 6 is susceptible of a geometrical interpretation, which we shall illustrate for the formula $(A-B-G+H)$ for "individual deviations from estimation." As explained earlier, the estimates for Method C are formed by superimposing a variation common to all the individuals upon the individual straight lines. The variation common to all the individuals is the deviation of the trial means from the group least square straight line. This latter is given by $(G-H)$. The individual lines are given by (B). Hence the estimates are given by $(B+G-H)$, and the deviations from estimation by $(A-B-G+H)$.

It is sometimes convenient to specify a certain combination of the sub-tables of Table 5 by what we may call a *code*, which indicates the separate items positionally rather than by letter. To illustrate, the code equivalent to the formula $A-B-G+H$ is $\begin{matrix}+&-&0\\0&0&0\\-&+&0\end{matrix}$. The plus and minus signs and zeros indicate which blocks of the array $\begin{matrix}A&B&C\\D&E&F\\G&H&J\end{matrix}$ are to be employed, and with what sign. The plus and minus signs correspond positionally with the blocks of the array to which they are attached, while the zeros indicate the blocks that are ignored.

Experimental Examples

The three following examples illustrate some of the commonest types of situation encountered in practise. Since we will be interested only in the tests for trend, the "between individual means" term will be omitted.

a. *Training with the Ball-pipe.* The Ball-pipe is a test of motor coordination, and the following analysis is from a period during which the

8 subjects were being trained in the test. The 5 trials were on April 29, 30, May 1, 3 and 5, 1943. For a detailed description of the experiment, the subjects and the nature of the test, see (1). In this example, $n=8$, $k=5$, and the values of x were taken to be -2, -1, 0, 1, 2. Thus x is a measure of the amount of practise, rather than of time. We are here interested in determining whether there is a significant training effect, and it may be to our advantage to admit the possibility of non-linearity in the individual graphs, as well as differences in slope from individual to individual. Thus we shall use the third method of estimation. The analysis is given in Table 7.

TABLE 7

TRAINING WITH THE BALL-PIPE

Source of variation	*Sum of squares*	*df*	*Mean square*
Individual deviations from estimation	106.0	21	5.048
Group deviations from linearity	17.6	3	5.867
Between individual slopes	59.15	7	8.450
Group slope	42.05	1	42.050

The ratio $F=5.867/5.048=1.162$ shows that the group graph does not deviate significantly from linearity; there is no consistent fluctuation apart from a linear trend. The ratio $F=42.050/5.048=8.330$ shows that the slope of the group graph is significantly different from zero. The means of the 5 trials are: 65.2, 64.2, 65.1, 63.8, 61.9; the analysis thus shows that the individuals are tending to lose the level of training which they had on April 29, which was the eleventh trial for these individuals.

If we had used Method *B* for our estimates, we would have obtained an analysis which would differ from that in the table only in the fact that the first two components would be combined or *pooled* to obtain a single term, namely, "individual deviations from linearity," with a mean square of 5.150, and 24 *df*. Using this value as an estimate of the error variance, we would have tested the group slope term, and it would be found to be significant, as before. The above analysis, under Method *C*, gives us a justification for pooling the first two terms, in that they are not significantly different, and thus provides a justification for the use of Method B. In fact, we may say that the analysis shows that the individual graphs are satisfactorily represented by straight lines; in so representing them we introduce an error no larger than the random variation of the experiment.

b. *Effect of experimental stress.* Consider as a second example the results of 3 trials of an intellective test, namely, the "Flags" subtest of the Factor Battery (2), derived from the Primary Mental Abilities battery. The 3 trials are 24 hours apart, and were obtained with 12

experimental subjects during a period of 65 hours without sleep. The first trial was in the evening before the first sleepless night, after 12 hours of wakefulness, while the two later trials were, respectively, after 36 and 60 hours of wakefulness. Since the trials were equally spaced, we may use for x the values -1, 0 and 1. The question to be answered here is, do the individuals show a tendency to deteriorate in their ability to perform this test as a result of prolonged wakefulness? The trial means are 55.6, 49.8 and 48.4. Is this drop significant? If we proceed under Method *C*, we obtain the analysis of Table 8.

TABLE 8

EFFECT OF EXPERIMENTAL STRESS

Source of variation	*Sum of squares*	*df*	*Mean square*
Individual deviations from estimation	83.445	11	7.586
Group deviations from linearity	37.555	1	37.555
Between individual slopes	368.833	11	33.530
Group slope	308.167	1	308.167

The ratio $F = 37.555/7.586 = 4.95$ is barely past the 5 per cent point, so that there is considerable doubt about the existence of a systematic fluctuation apart from linear trend. On the other hand, the ratio $F = 308.167/7.586 = 40.62$ indicates that we have a significant linear trend for the whole group. We conclude that this intellective function shows significant deterioration under prolonged wakefulness.

c. *Incidental trend in an experiment.* The third example deals with data from an experiment with 12 normal individuals, to determine the extent of the effect of a meal upon certain electrocardiographic measurements. Six trials were obtained upon 12 individuals, and for each trial electrocardiograms were obtained before and after a meal. Of the several electrocardiographic functions that were analysed only one will be considered here, namely, the function known as "sigma T", the sum of the absolute values of the T-waves in all three leads. The mean differences in this function, due to the effect of the meal were, for the six trials: -1.14, -2.63, -3.19, -2.48, -2.62, -2.01. Quite apart from the question as to whether the mean difference for all 72 pairs of observations is significant, we may ask whether the differences show a significant trend from trial to trial. The analysis was carried out under Method C, with the results given in Table 9.

In this case there is no significant group slope, but the group deviations from linearity are highly significant. The interpretation is that the individuals have shown no progressive change in their response to the meal, but there have been highly significant *fluctuations* about the mean value. The fact that these fluctuations are significant implies that

all the 12 individuals have, to some extent, participated in them, and thus they represent a common experimental condition, perhaps due to the differences between the 6 meals. The "group deviations from

TABLE 9

INCIDENTAL TREND IN AN EXPERIMENT

Source of variation	*Sum of squares*	*df*	*Mean square*
Individual deviations from estimation	24.391	44	.554
Group deviations from linearity	14.197	4	3.549
Between individual slopes	4.825	11	.439
Group slope	.515	1	.5

linearity" term will frequently represent this type of variation, and may then be taken as indicative of some common experimental condition.

COMPARISON OF GROUP TRENDS

The methods developed above can be extended to provide a means of comparing group trends. While it is possible to describe the method in geometrical terms such as were used in developing Methods A, B and C for a single group, it is probably simpler to use a numerical approach, based on the sub-tables of Table 5.

We now suppose that the 4 individuals of Table 1 fall into two groups, as indicated in that table, and we are interested in determining whether these groups show different trends. From now on we use the term "group" to refer to these sub-groups, while the adjective "overall" will be used to refer to all four individuals collectively. Thus, in Table 5, Table B consists of individual linear estimates, Table E of group linear estimates, and Table H of overall linear estimates.

The 8 components of the analysis are given in detail in Table 10. Each is obtained from a certain combination of the sub-tables of Table 5, and these combinations have been indicated both by letters and by "codes" such as were described earlier.

The actual tables of components have been exhibited, since these reflect many of the important features of the analysis. In addition to the 8 components of variability, a table of the overall mean has been included in the lower right, in order to make it a simple matter to verify that the 20 original items of Table 1 may be obtained by adding corresponding items of the 9 sub-tables of Table 10. The sum of squares (ss) and df are given below each sub-table, along with the mean square ($ms = ss/df$), so that it can readily be verified that any sum of squares can be obtained either by adding the squares of the items in the table or by applying the code to the ss values given in Table 5.

The most important components are the four in the upper left, and these require some explanation.

TABLE 10

Individual deviations from estimation $(A-B-D+E)$					*Between individual slopes* $(B-C-E+F)$					*Between individual means* $(C-F)$				
		+	−	0			0	+	−			0	0	+
		−	+	0			0	−	+			0	0	−
		0	0	0			0	0	0			0	0	0
41	−22	−55	12	24	−16	−8	0	8	16	−35	−35	−35	−35	−35
−41	22	55	−12	−24	16	8	0	−8	−16	35	35	35	35	35
−3	26	−5	−56	38	−12	−6	0	6	12	5	5	5	5	5
3	−26	5	56	−38	12	6	0	−6	−12	−5	−5	−5	−5	−5
ss = 22,400 df = 6 ms = 3,733					ss = 2,000 df = 2 ms = 1,000					ss = 12,500 df = 2 ms = 6,250				
Group deviations from estimation $(D-E-G+H)$					*Between group slopes* $(E-F-H+J)$					*Between group means* $(F-J)$				
		0	0	0			0	0	0			0	0	0
		+	−	0			0	+	−			0	0	+
		−	+	0			0	−	+			0	0	−
5	−35	35	15	−20	10	5	0	−5	−10	−5	−5	−5	−5	−5
5	−35	35	15	−20	10	5	0	−5	−10	−5	−5	−5	−5	−5
−5	35	−35	−15	20	−10	−5	0	5	10	5	5	5	5	5
−5	35	−35	−15	20	−10	−5	0	5	10	5	5	5	5	5
ss = 12,400 df = 3 ms = 4,133					ss = 1,000 df = 1 ms = 1,000					ss = 500 df = 1 ms = 500				
Overall deviations from linearity $(G-H)$					*Overall slope* $(H-J)$					*Overall mean* (J)				
		0	0	0			0	0	0			0	0	0
		0	0	0			0	0	0			0	0	0
		+	−	0			0	+	−			0	0	+
14	−3	−40	33	−4	−14	−7	0	7	14	300	300	300	300	300
14	−3	−40	33	−4	−14	−7	0	7	14	300	300	300	300	300
14	−3	−40	33	−4	−14	−7	0	7	14	300	300	300	300	300
14	−3	−40	33	−4	−14	−7	0	7	14	300	300	300	300	300
ss = 11,640 df = 3 ms = 3,880					ss = 1,960 df = 1 ms = 1,960					ss = 1,800,000 df = 1				

(a) Individual deviations from estimation, $(A-B-D+E)$

This is the fundamental error term of the analysis. It consists of the deviations of the original values (A) from the estimated values $(B+D-E)$. Since $(D-E)$ gives the deviations of the group trial values from linearity, while (B) represents the individual least square straight lines, the estimate consists of the group deviations from linearity super-

imposed upon the individual linear estimates. The error term is simply a measure of the failure of the individual values to conform to this estimate.

(b) Group deviations from estimation, $(D-E-G+H)$

Here the estimated values are given by $(E+G-H)$. $(G-H)$ gives the deviations of the overall trend from linearity while (E) gives the group linear estimate. Thus the estimate $(E+G-H)$ consists of the overall deviations from linearity superimposed on the group linear estimates. If the mean square for $(D-E-G+H)$ were zero, it would mean that the group deviations from linearity were identical with the overall deviations from linearity. If, on the other hand, this mean square is significantly greater than error, it implies that the group deviations from linearity are significantly greater than the overall deviations from linearity, which in turn implies that the group deviations from linearity are significantly different *from each other*. This term thus enables us to recognise a non-linear component of variation which is different for the different groups.

(c) Between individual slopes, $(B-C-E+F)$

This term is simply a measure of whether the individual slopes differ significantly from each other within the groups.

(d) Between group slopes, $(E-F-H+J)$

This term provides a test of whether the two groups differ significantly in slope. It is this term which will generally determine whether the groups differ as to trend.

In Table 10 the fundamental error term is 3,733. When this mean square is compared with the seven others, it is found that none of them is significantly greater than the error term, and it is concluded that only one source of variation is present in the table. This is the conclusion that is to be anticipated in view of the fact that the original table was constructed from random numbers.

In the practical method of dealing with this type of example we work directly from Table 1, and construct the 9-block Table 11.

Note that the 9 blocks of Table 11 correspond to the 9 blocks of figures in Table 1. Six of the sums of squares in Table 11 have already been computed and used in the analyses under Methods A, B and C. The remaining calculations should be clear with the understanding that $\Sigma_1 P$ and $\Sigma_2 P$ represent, respectively, the sum of the values of P for the first and second group, namely, 40 and 240.

We may conveniently indicate, by a code similar to that of Table 10, how the sums of squares and *df* given in Table 10 may be derived from Table 11. The necessary codes are given in Table 12.

When these codes are applied to the 9 blocks of Table 11, the result-

ing mean squares and *df* will be found to agree with those in Table 10.

Let us now extend the above results to the general case in which

TABLE 11

$\Sigma Y^2=1{,}864{,}400$ $df=20$	$\Sigma I^2/5=1{,}813{,}000$ $df=4$	$\Sigma P^2/10=4{,}960$ $df=4$
$[\Sigma T_1^2+\Sigma T_2^2]/2$ $=1{,}827{,}500$ $df=10$	$[(\Sigma T_1)^2+(\Sigma T_2)^2]/10$ $=1{,}800{,}500$ $df=2$	$[(\Sigma_1 P)^2+(\Sigma_2 P)^2]/20$ $=2{,}960$ $df=2$
$\Sigma T^2/4=1{,}813{,}600$ $df=5$	$GT^2/20=1{,}800{,}000$ $df=1$	$(\Sigma P)^2/40=1{,}960$ $df=1$

there are p groups with n_1 individuals in the first, n_2 in the second, $\cdots$, and n_p in the last. Then, if n is the total number of individuals,

$$n=n_1+n_2+\cdots+n_p.$$

Let k be the number of trials. The various summations should be arranged as in Table 1 (although the means need not be calculated), so

TABLE 12

Individual deviations from estimation + − − − + + 0 0 0	*Between individual slopes* 0 0 + 0 0 − 0 0 0	*Between individual means* 0 + 0 0 − 0 0 0 0
Group deviations from estimation 0 0 0 + − − − + +	*Between group slopes* 0 0 0 0 0 + 0 0 −	*Between group means* 0 0 0 0 + 0 0 − 0
Overall deviations from linearity 0 0 0 0 0 0 + − −	*Overall slope* 0 0 0 0 0 0 0 0 +	

that we have 9 blocks of numbers, from each of which we will obtain a sum of squares. Let $T_1, T_2, \cdots, T_p$ represent the trial totals for the groups, let $\Sigma T_1, \Sigma T_2, \cdots, \Sigma T_p$ represent the group totals, and let $\Sigma_1 P, \Sigma_2 P, \cdots, \Sigma_p P$ represent the subtotals of P for each of the

groups. A table of sums of squares and *df* may then be constructed as shown in Table 13.

TABLE 13

$ss = \Sigma Y^2$ $df = nk$	$ss = \Sigma I^2/k$ $df = n$	$ss = \Sigma P^2/\Sigma x^2$ $df = n$
$ss = \frac{\Sigma T_1^2}{n_1} + \frac{\Sigma T_2^2}{n_2} + \cdots + \frac{\Sigma T_p^2}{n_p}$ $df = pk$	$ss = \frac{(\Sigma T_1)^2}{kn_1} + \frac{(\Sigma T_2)^2}{kn_2} + \cdots + \frac{(\Sigma T_p)^2}{kn_p}$ $df = p$	$ss = \frac{(\Sigma_1 P)^2}{n_1 \Sigma x^2} + \cdots + \frac{(\Sigma_p P)^2}{n_p \Sigma x^2}$ $df = p$
$ss = \Sigma T^2/n$ $df = k$	$ss = GT^2/nk$ $df = 1$	$ss = (\Sigma P)^2/n\Sigma x^2$ $df = 1$

Note that in forming each of the totals in the second row, we form a sum of squares for each group separately and then add. There is considerable simplification for the case in which all the groups have the same number of individuals, since all the denominators of the fractions in each block would then be the same.

TABLE 14

EFFECT OF THIAMIN DEFICIENCY ON FLAGS TEST

x	−12.6	−6.6	.4	7.4	11.4	I	P
G	60	61	59	52	53	285	−146.0
Wi	62	67	65	54	65	313	− 56.8
S	69	68	64	68	59	328	−116.8
Wa	72	71	76	70	76	365	39.0
T	46	43	43	46	48	226	41.4
N	46	43	45	44	47	225	16.0
Jo	81	79	80	84	81	405	35.0
T_1	191	196	188	174	177	926	−319.6
T_2	245	236	244	244	252	1221	131.4
T	436	432	432	418	429	2147	−188.2

To complete the analysis, we have only to apply the codes given in Table 12 to the above set of 9 blocks.

An experimental example which may be treated by the above method is given in Table 14. The test is the same "Flags" test that was referred to above, in the example on the effect of an experimental stress. During the period of the five trials, which covered 24 days, the 7 subjects were on a basic diet which provided only about .03 mg of

thiamin per day. The first group of 3 men received no supplementation, and hence may be referred to as the deficient group. The second group received a supplement of 1.50 mg of thiamin per day, and will be referred to as the supplemented group. These groups will be denoted by G_1 and G_2 respectively.

In Table 14 the original data and the various types of sums have been separated into 9 blocks, as in Table 1. Table 15 shows how we obtain the sum of squares and *df* for each of these blocks, following the outline in Table 13.

TABLE 15

$ss = \Sigma Y^2 = 137{,}465.0$ $df = 35$	$ss = \Sigma I^2/5 = 137{,}145.8$ $df = 7$	$ss = \Sigma P^2/387.2 = 110.8$ $df = 7$
G_1: $\Sigma T_1^2/3 = 57{,}282.0$ G_2: $\Sigma T_2^2/4 = 74{,}574.2$ $ss = 131{,}856.2$ $df = 10$	$(\Sigma T_1)^2/15 = 57{,}165.1$ $(\Sigma T_2)^2/20 = 74{,}542.1$ $ss = 131{,}707.2$ $df = 2$	$(\Sigma_1 P)^2/1{,}161.6 = 87.9$ $(\Sigma_2 P)^2/1{,}548.8 = 11.1$ $ss = 99.0$ $df = 2$
$ss = \Sigma T^2/7 = 131{,}729.9$ $df = 5$	$ss = GT^2/35 = 131{,}703.1$ $df = 1$	$ss = (\Sigma P)^2/2{,}710.4 = 13.1$ $df = 1$

The codes of Table 12 are now used to complete the analysis, as exhibited in Table 16.

TABLE 16

Individual deviations from estimation $ss = 158.4$ $df = 15$ $ms = 10.6$	*Between individual slopes* $ss = 11.8$ $df = 5$ $ms = 2.4$	*Between individual means* $ss = 5{,}438.6$ $df = 5$ $ms = 1{,}087.7$
Group deviations from estimation $ss = 36.3$ $df = 3$ $ms = 12.1$	*Between group slopes* $ss = 85.9$ $df = 1$ $ms = 85.9$	*Between group means* $ss = 4.1$ $df = 1$ $ms = 4.1$
Overall deviations from linearity $ss = 13.7$ $df = 3$ $ms = 4.6$	*Overall slope* $ss = 13.1$ $df = 1$ $ms = 13.1$	

The fundamental error term is "individual deviations from estimation," with a mean square of 10.6. The fact that neither of the terms below it is significantly greater than error indicates that neither in the overall group nor in the sub-groups is there any systematic deviation from linearity. This has the consequence that we could have used a purely linear method of analysis, such as was developed for a single group in Method B. In this case, the three terms in the first column would be pooled to obtain a mean square of 9.93 with 21 *df*, which would be termed the "deviations from linearity" mean square.

"Between individual slopes" is of the same magnitude as error, indicating that the slopes of the individual least square straight lines are not significantly different within each group. The critical term in the analysis is "between group slopes," whose value is 85.9. This is highly significantly greater than error, whether we take the error term to be 10.6 with 15 *df*, or 9.93 with 21 *df*. The implication is that if we fit a least square straight line to the total body of data for Group 1, and another such line to the data of Group 2, then the slopes of these two straight lines are significantly different. The mean values for the two groups are:

Group 1 (Deficient):	63.7	65.3	62.7	58.0	59.0
Group 2 (Supplemented):	61.2	59.0	61.0	61.0	63.0

It is clear that the deficient group has deteriorated, while the supplemented group has maintained a plateau.

The remaining terms, in the third column of the table, are of no interest to us, since they reflect the initial levels of the experiment. We may, however, remark that the mean square term "between group means" $=4.1$ implies that the groups were fairly well matched initially.

Summary

In treating the results of experiments in which a number of individuals are measured simultaneously on a number of trials, methods are presented for determining whether the group as a whole has deviated from plateau performance. Where trend is present, the methods enable one to distinguish linear trend from a fluctuation in which the whole group participates. An extension of these methods permits the comparison of the trends of two or more groups. Geometrical and arithmetical illustrations are used to bring out the underlying structure of the analysis of variance in the several methods.

BIBLIOGRAPHY

1. BROZEK, J., GUETZKOW, H., MICKELSEN, O., & KEYS, A. Motor performance of normal young men maintained on restricted intakes of Vitamin B Complex. *J. appl. Psychol.*, 1946, **30**, 359–379.
2. GUETZKOW, H. & BROZEK, J. Intellectual functions with restricted intakes of B Complex Vitamins. *Amer. J. Psychol.*, 1946, **59**, 358–381.
3. JACKSON, R. W. B. Reliability of mental tests. *Brit. J. Psychol.*, 1939, **29**, 267–287.
4. JACKSON, R. W. B. *Application of the analysis of variance and covariance method to educational problems.* Bull. No. 11, Dept. of Educational Research, Univ. of Toronto, 1940.
5. JACKSON, R. W. B. *Studies on the reliability of tests.* Bull. No. 12, Dept. of Educational Research, Univ. of Toronto, 1941.
6. SNEDECOR, G. W. *Statistical methods.* (4th Ed.) Ames, Iowa: Collegiate Press, 1946.

A NOTE ON GRANT'S "NEW STATISTICAL CRITERIA FOR LEARNING AND PROBLEM SOLUTION"

IRVIN L. CHILD
Yale University

In Grant's excellent article, *New statistical criteria for learning and problem solution in experiments involving repeated trials*, recently published in this journal (2), there appears an error of omission which has consequences of some importance and which therefore merits attention. The error consists of failure to make a correction for continuity in applying the normal curve and the χ^2 approximations. The probable consequences are (1) undue discouragement of the use of these approximations; (2) when the approximations are used, an easily avoidable exaggeration of the significance of the findings.

This omission appears in Grant's treatment of the criterion of total number or proportion of correct responses. The omission and its consequences may well be considered with reference to the illustrative problem which Grant presents (2, pp. 275f.). A rat is reported to have responded correctly in 19 out of 25 successive trials. The problem is to find the probability value expressing the significance of the deviation from a hypothetical 12.5 correct response out of 25. Precise calculation of the probability gives a value of .0073. Grant's application of the normal curve approximation gives a normal deviate of 2.4 and a probability of .0047, with the "very considerable error of approximation" of $-.0026$. Application of the χ^2 approximation in the manner suggested by Grant gives a χ^2 of 6.76 and an identical probability of .0047. This probability value, not given by Grant, was determined by treating $\chi = 2.6$ as a normal deviate, a method which is precise but applicable only with a single degree of freedom. (The probability cited by Grant as between .01 and .005—the exact value is .0094—is not comparable; it refers to the probability of obtaining so large a deviation *in either direction* from theoretical frequencies.) The error of approximation is of course identical with that found in direct application of the normal curve.

Actually, these errors have been considerably exaggerated by failure to make the correction for continuity. As Yates has pointed out (6), it is necessary in applying a continuous function (such as the normal curve) to data which are by their nature discrete (as are frequencies), to give the discrete numbers a meaning on the continuous scale. The meaning required by the conditions of these problems is, for each integer

representing a frequency, an interval from one-half unit below to one-half unit above the integer. This meaning requires that a reduction of one-half point be made in the absolute magnitude of the deviation of each observed frequency from the corresponding theoretical frequency. This correction is widely known as Yates' correction for continuity.*

In Grant's illustrative problem, the correction is readily applied to the normal curve approximation by simply reducing the difference from 6.5 to 6.0. The result is a normal deviate of 2.4 and a probability of .0082. The true error of approximation is thus +.0009, about one-third of that cited by Grant.

In the case of the χ^2 approximation, the correction is equally readily made by one who is following the general procedure for the calculation of χ^2, as described in most statistical textbooks. All that is necessary is to reduce by one-half unit the absolute magnitude of each difference between an observed and a theoretical frequency, before proceeding to further calculation. If the equations given by Grant are used they may be so altered as to effect the correction. His equation [4] becomes:

$$\chi^2 = \frac{\left(m - \frac{p}{q}s \pm \frac{0.5}{q}\right)^2}{\frac{p}{q}n},$$

where $0.5/q$ is to be added if the sum of the two preceding terms is negative, and to be subtracted if that sum is positive. The corrected form of Grant's equation [5], for the special case where $p=q=0.5$, is the following:

$$\chi^2 = \frac{(m - s \pm 1)^2}{m + s},$$

where again the addition or subtraction of the third term within the parenthesis depends upon whether the sum of the preceding terms is negative or positive respectively. In both these equations, in other words, the sign of the third term is so taken that the effect will be to reduce the absolute magnitude of the quantity within the parentheses.

When χ^2 is calculated for Grant's illustrative problem, by any of these methods which make the correction for continuity, its value is

* For other accounts of Yates' correction for continuity, see Goulden (3, pp. 101–108) and Rider (5, pp. 112–115). Both of these authors, as well as Yates (6) give examples of the amount of error introduced through failure to apply it in types of problems different from those being discussed here.

found to be 5.76. The corresponding probability value is .0082, and the error of approximation is +.0009, just as in the direct application of the normal curve approximation.

The following statement by Grant, finally, must be rejected: "In general [4] and [5] will give more accurate approximations than [3] when n is small." It may readily be demonstrated that equations [3] and [4] are mathematically equivalent except that [3] gives the value of x/σ while [4] gives the value of χ^2, which is simply the square of x/σ.* Either value can be precisely determined through the use of either equation with identical results. The probabilities arrived at, if correctly determined, are also identical. (For a treatment of the relationship between χ^2 and the normal curve, see, for example, Peters and Van Voorhis, (**4**, ch. XIV).) It is apparent, therefore, that neither method can have any advantage over the other except convenience.

It is hoped that these comments may encourage the use of the extremely valuable criteria put forth by Grant by calling attention to the rather satisfactory approximation possible through the time-saving use of the normal curve or χ^2. But the user will certainly do well to bear in mind the inaccuracies which may result when n is small or either p or q approaches unity, and to observe limits such as Grant suggests for the use of these short-cuts.† (The limits, however, should be the same for both techniques, since the two techniques are essentially the same.)

Addendum

After this note was written, Dr. Grant called my attention to the fact that the note may be misleading because it seems to imply that the use of Yates' correction always yields directly a better approximation to the exact probability than would be obtained without the correction. This is sometimes far from true in parts of the distribution for cases where the theoretical probability differs from 0.5, a point that Yates illustrates in his Table II (**6**, p. 223). Fortunately, however, it is not necessary to rely upon a guess as to which method gives the better approximation in a particular case. Yates has developed a technique by which the corrected χ^2 may be evaluated by reference to a table which takes account of the effects of asymmetry, permitting a test of significance which appears to be quite accurate enough for most purposes.

* What is said of Grant's equation [4] is also true of his equation [5] for the special case where it is applicable.

† An exception to this statement about limits can be made, however, if χ^2 is used and evaluated by the technique referred to here in the addendum. This technique developed by Yates appears to be quite applicable even when n is small or p or q approaches unity.

The technique is presented in Yates' original article (Table III on p. 228, and the accompanying text). It is, however, more accessible in Fisher and Yates' *Statistical Tables* (1), where it is presented in Table VIII and in the corresponding text in the introduction. Because of the considerable errors that may result from failure to make these corrections, use of this technique should probably be standard practice with 2×2 contingency tables (where a theoretical probability of 0.5 is exceptional), and with those problems of the sort posed by Grant in which the theoretical probability is other than 0.5.

BIBLIOGRAPHY

1. FISHER, R. A. & YATES, F. 2nd Ed., Rev. *Statistical tables for biological, agricultural and medical research.* London: Oliver & Boyd, 1943. (The first (1938) edition also contains the table referred to.)

2. GRANT, D. A. New statistical criteria for learning and problem solution in experiments involving repeated trials. *Psychol. Bull.*, 1946, **43**, 272–282.

3. GOULDEN, C. H. *Methods of statistical analysis.* New York: Wiley, 1939,

4. PETERS, C. C. & VAN VOORHIS, W. R. *Statistical procedures and their mathematical bases.* New York: McGraw-Hill 1940.

5. RIDER, P. R. *An introduction to modern statistical methods.* New York: Wiley, 1939.

6. YATES, R. Contingency tables involving small numbers and the χ^2 test. *Suppl. J. roy. stat. Soc.*, 1934, **1**, 217–235.

"OPINION-ATTITUDE METHODOLOGY" AND THE POLLS — A REJOINDER*

LEO P. CRESPI
Princeton University

In the July issue of this journal, Professor Quinn McNemar of Stanford University brought forth an 85-page assessment of attitude and public opinion methodology (**5**). Such a project gives a writer an opportunity to be of real service to the infant discipline of public opinion measurement by a fair weighing of the merits and demerits of its present procedures. Perusal of McNemar's critique quickly makes it evident that either public opinion measurement is almost totally lacking in merits, or that the critic employed not the scales of even-handed appraisal but a microscope, to occupy himself almost exclusively with the magnification of demerits. If the first alternative is true, it is singular to find so many groups whose functioning is based upon dependable information—government administrators, politicians, editors, industrialists, and labor leaders—uncomplainingly paying out larger and larger sums for public opinion analysis. But what about the second alternative? It is the present writer's purpose to present a few of the considerations pointing to the existence of an anti-polling bias which seriously impugns the critique.

The evidence of bias is, primarily, repeated errors in McNemar's discussion which are unfailingly in the same direction—namely anti-polling—and hence, difficult to account for on the basis of chance oversights alone. For convincing illustration one need go no farther than the second paragraph on page 314 of McNemar's paper which is quoted below:

> The only information on the reliability of opinion questions is in Cantril's *Gauging Public Opinion*, and this pertains only to one question: 'Do you think Roosevelt is doing a good job, only a fair job, or a bad job running the country?' For interview-reinterview, with a three weeks interval, the percentage of identical responses given by 286 persons was 79, which for the given situation tends to approximate a correlation coefficient of .90. This question is said to be one of the most stable, hence it may be inferred that other questions would yield lower reliabilities. By the same method 87% of the same group gave consistent answers to a question concerning whom they voted for in the last presidential election, and information given about car ownership (car or nor car) agreed only 86% of the time. Evidently the factual type of question is none too reliable, so one wonders just what the story is for opinion questions.

* Dr. Herbert S. Conrad and Dr. Richard Centers have been kind enough to read this manuscript. They agree substantially with the opinions expressed.

This passage, it should be noted, is the backbone of one of McNemar's major criticisms of the polls—their asserted deficiency in reliability as McNemar defines the term. The passage is disparaging, which, however, is not the point at issue. Is it justifiably disparaging? The answer must be no as from first to last the passage is permeated with errors anti-polling in their effects. In the opening line the contention is made that, "The only information on the reliability of opinion questions is in Cantril's *Gauging Public Opinion* . . . " This statement is an error and an unfortunate one in terms of what McNemar might have learned about the polls from the articles he overlooked. Without endeavoring to be complete the present writer can cite three studies: (Dodd (**2**), Jenkins (**3**), King (**4**), in the literature which approach in detail the reliability of the individual's responses to interview questions.

These studies do not exhaust the problem, but if McNemar had seen them he would have appreciated the injustice of his remark, that, "This aspect of reliability [individual reliability] has been practically ignored by the users of the single question technique, although they call their work 'scientific' " (p. 314). The three articles would together have invalidated the accusation of ignoring the problem. Dodd's article, in addition, would have exposed the erroneous assumption that preoccupation with individual reliability is necessary to all polling that would be *scientific.* Dodd repeatedly demonstrated under particularly handicapping conditions of polling in war zones that "almost perfect" group or as he calls it *plurel* reliability can be obtained with quite imperfect individual reliabilities—these latter individual fluctuations consistently cancelling each other out. So if, as is frequently the case, the poller is only interested in reliably characterizing the opinion of the group as a whole, the reliability of individual responses can be ignored without impugning the scientific quality of the work. The poller need only be assured of adequate group reliability.

One more thesis McNemar might have revised had these studies not been overlooked, this is his prediction upon page 325—"It is the writer's belief that the reliability for single questions will be found to be low; if measured by some coefficient equivalent to product moment correlation, we estimate that typical reliabilities will be in the .50's or .60's, sometimes lower and occasionally higher." All three of the studies that McNemar failed to consider definitely suggest that the reliabilities (for individuals) of public opinion poll questions are substantially higher than this prediction. Interested readers are invited to check the facts here for themselves. However, to conclusively settle the question the present writer has initiated a study of the reliability of typical poll questions.

Returning to the paragraph that has been quoted from McNemar's critique, let us now focus our attention upon the last two sentences. If one turns to the study McNemar cites, that presented by Mosteller in Cantril's *Gauging Public Opinion* (1, chap. 7), it is found that the key concern in the discussion of reliability is to compare the results obtained by two modes of measurement—interview-reinterview by the same interview *vs.* by two different interviewers. For the question about car ownership to which McNemar refers, quite different reliability results were reported for the two methods of measurement—with different interviewers, 86 per cent agreement upon repetition of the question; with the same interviewer, 96.5 per cent agreement. In reporting these results a critic could have

1. cited only the higher reliability index and based his conclusions thereon —this could suggest a pro-poll bias,
2. cited both figures, decide on an appropriate measure, and draw his conclusions therefrom—this is the fair and neutral tack.
3. cited only the lower figure and base his conclusions thereon—this is the tack McNemar takes.*

But questionable selection is only the beginning. Not only is the less flattering figure alone mentioned, but in addition, it is erroneously represented as obtained by the method which in fact gave the higher estimate of reliability. That is to say, the "same method" as the Roosevelt question (namely same interviewers) gave 96.5 per cent reliability agreement, not the 86 per cent McNemar states. So there is not only a singular omission of data (keeping in mind that McNemar is all the while vigorously bewailing the paucity of data on the point) but also an outright misstatement. That the effect of such shortcomings is serious is easily appreciated by merely substituting the correct value of 96.5 for McNemar's 86 per cent and noting what happens to his final point.

To complete the itemization of errors to be found in the one short paragraph that was quoted, it must be also mentioned that the 87 per cent reliability index which McNemar cites for the question, whom one voted for in the last election, (a) was not obtained as represented "By the same method . . . " as in the Roosevelt question which preceded it, and (b) was not obtained "of the same group."

We have gone to some length in this discussion of reliability not to construct a catalogue of errors—which would simply be following in

* To observe the difference between these latter two types of presentation, compare McNemar's discussion of Mosteller's findings with the presentation of the same material by Newman, Bobbitt, and Cameron, (6).

McNemar's footsteps—but to demonstrate that this "appraisal, by an outsider"—"someone without prior or vested interest" (p. 289) is not the complete and impartial assessment of the polls that it represents itself to be. An appreciation of this fact impresses the writer as being of greater importance in putting the critique in proper perspective than presentation of detailed replies to the many dubious contentions. These latter more detailed rejoinders will come later, and if the writer is any judge, in a deluge.*

Without going into the same detail, another sample paragraph may be advantageously scrutinized, which contains two of McNemar's criticisms of the questions employed in polling, namely "*understandability of the question and/or the issues involved*" (p. 318) [italics McNemar's]. As the basis for the first, McNemar singles out a question from a recent study of the Office of Public Opinion Research, which we may quote here:

> For handling domestic problems like unemployment, the converting of war plants to peacetime use, or the demobilization of soldiers—do you think the government should set up a central agency *now* with full authority to make plans and with full authority to carry out these plans as soon as the war is over?" (p. 318).

McNemar then goes on to make his point—"The writer of that question certainly has an ivory tower notion regarding the intellectual span of the average adult." That this question which McNemar has carefully selected to clothe his point is open to legitimate criticism is undeniable, but that it is quite unrepresentative is equally so. McNemar's criticism, however, is over-drawn as he presents it. He neglects entirely to take the fact into account that questions, no less than statements, must not be torn from context. The particular query cited appeared late in a ballot of which it formed a meaningful and timely part, and which illuminated the issue. Further, how can one be certain that the question *was* too difficult for the average respondent, even if posed in isolation? Are pollers to accept by fiat what is the intellectual span of the average adult? Occasionally the purpose of the polls in including a difficult question is precisely to find out just what is the intellectual understanding of the population.

It is McNemar's second criticism of poll questions which is the more significant in its implications. This is his censure of certain questions because of the kind of *issue* they present to the people. As McNemar presents this criticism it seems to point to a philosophy which could hardly make its holder other than inhospitable to the polls. McNemar

* See for example, the lengthy rejoinder by Dr. H. S. Conrad in this issue.

states, "Questions are also being asked about the postwar world, the future peace, and other rather abstract issues, . . ." the answers to which McNemar contends become " . . . illustrations of 'a consensus of worthless opinion is a worthless consensus of opinion' " (p. 318). In short, McNemar holds, it appears, that the opinions of the people upon such subjects as the postwar world and the future peace are "worthless." This is a remarkable thesis. Does McNemar realize such a criticism of polling parts company with Democracy also? It is the essence of the democratic faith and a working principle of pollers that the opinion of the people upon issues so fundamental to their welfare is of *supreme* importance, this regardless of the fact that they may not be experts by virtue of intellect or experience. So if, as it would seem, the polls are now to be charged with being democratic in inspiration, they can only plead guilty.

McNemar confesses to "no hesitance" in making the proposal that "*single question opinion gauging be discarded in favor of opinion measurement by attitude scales*" (p. 327) [italics McNemar's]. Such a stand cannot fail to impress most students of polling, even those most critical of its present weaknesses, as throwing out the baby with the bath. The additional thought will occur to some that such action is not so much an oversight as great affection for another infant. McNemar endeavors to forestall the obvious rejoinder to his position with the following statement: "If it is objected that a given problem does not justify the extra work entailed in scale construction and the added schedule administration time, there is reason to suspect that the project isn't of much value from a rigorous scientific viewpoint" (p. 328). Much that is questionable can hide behind this persuasive-sounding claim. McNemar's unhesitating suggestion would seem to imply that the bulk of AIPO, *Fortune* and NORC material is of little value. Taken together with his repeated insistence upon the utmost rigor in investigation McNemar's position reveals a definite misconception. He makes the perfectionist's error, possibly an occupational disease with methodologists, of single-trackedly seeking to maximize precision in scientific experiments. Is not the proper function instead to maximize results per unit effort? It is easily possible that the increase in information to be obtained with the substitution of expensive and laborious scales for the simple "single question" is relatively small and far outweighed by the increase in effort. Furthermore, if people's interviewing tolerance is already being strained by using single questions upon issues, what will be the consequences of employing lengthy scales?

McNemar's emphasis upon methodological and statistical refine-

ment would seem to excommunicate a whole new area of modern statistical development which is concerned with the fashioning of relatively imprecise statistics which are, however, of remarkable efficiency in terms of information obtainable per unit cost in time and money. Whatever may be the inclination of the methodologist, the concrete researcher knows well that his yardstick of success in experimentation must be efficiency—the results per unit cost. Over-preoccupation with refinement can be just as sterile in science as in social life.

Shifting from Professor McNemar's specific discussion to his introductory more generalized remarks, does one find here any indications of anti-poll distortions? The present writer feels that the answer must be yes. A case in point may quoted, " . . . it is reasonably certain that the psychologists and sociologists are mainly responsible for the current status of attitude research (by scales), and that the journalists have contributed much to opinion research (by single questions)" (p. 292). The obvious implication is that psychologists and sociologists, i.e. "scientists," are mainly responsible for the current status of attitude research by scales, but not of opinion research by single questions. This is untrue. This statement disparages opinion research by the suggestion that it is conducted by journalists, i.e. laymen untrained in "scientific" method. The fact is that psychologists and sociologists (in academic halls and in commercial agencies) are also mainly responsible for the current status of opinion research by single questions. What *is* true is that the findings of opinion research have proved so informative in attacking the problems of everyday life that journalists have been motivated to make wide use of poll results in their writings. But this is a far cry from what McNemar's statement implies.

If McNemar's discussion of the reliability of the polls shows evidence of biased handling, his efforts to castigate the polls on the basis of validity go as far as plain contradiction. On page 315 he states (a) "One might have expected that Cantril's volume would have tackled the problem of opinion question validity but such is not the case." But what is the statement to be found immediately afterwards? (b) "The volume does touch on some of the factors which contribute to . . . invalidity, . . . "! So this critic, in statement (a), accuses Cantril of omitting any discussion of validity—which is unwarranted. He protects himself, however, by making statement (b)—which is certainly true, since the entire point of chapter 1 and, in part, chapter 2 of Cantril's volume is the validity of questions. What is the result? Three errors—an error of fact, an error of logic, and an error of unjustified condemnation. The only legitimate statement that McNemar could have made in

the present connection is that Cantril's remarks bearing upon the validity of poll questions are incomplete—which is hardly surprising in view of the fact that the entire book was explicitly introduced as being precisely that.

It appears evident, in conclusion, that the only way to avoid a place in Professor McNemar's syllabus of errors is to follow the old precept, "Them that does nothin', never gets into trouble." In the present critique this adage acquires a new twist—"Them that does the most, gets into the most trouble." For it seems that those who have done the most work in the field are subjected to the most criticism.

Those of McNemar's criticisms of the polls which bear up under examination are, on the whole, far from novel. They have already been as well, or better, expressed by pollers themselves. Failure to make this fact clear in McNemar's critique leads to the unwarranted impression that the pollers have been guilty of repression or suppression of criticism. The deficiencies in public opinion analysis, and certainly there are many, do not usually spring from ignorance or ineptness. Polling technique even in its initial crude state proved to be so useful in the world of public affairs that there has always been heavy pressure upon opinion researchers to spend their time obtaining concrete information by the available methods. During the war the necessity for information was so immense that methodology, in considerable part, had to mark time. First things first—and the government and other groups who were financing many of the polls were paying for quick answers to their questions, not for large scale methodological inquiries.

In the face of such pressures it is rather remarkable that so many methodological studies were published during the war and that a complete volume on the subject could be contributed by the Office of Public Opinion Research. To repeatedly single out for attack this book of my colleague because it does not answer all questions—and the studies, *after* they are read, reveal some of their own limitations—is not the mark of a reasonable critic. This is especially true when Cantril makes it quite clear in his preface to *Gauging Public Opinion* the spirit in which it was offered. To quote:

> Because the field is still in its infantile stages, the investigator today feels somewhat as must the explorers of the fifteenth and sixteenth centuries when they saw only bits of vast unknown areas whose precise boundaries and potentialities they knew they themselves would never fully comprehend. So many problems in the field call for some solution that the investigator hardly knows where to begin. As soon as he begins work on one problem, several other related and equally important problems loom ahead.
>
> In spite of the miscellaneous nature of the studies reported here, I have

tried to bring together for the first time in one volume some idea of the serious problems encountered in every phase of the polling operation. My hope is that the series of studies will advance the science of polling, show that it is more than a parlor game, and point up some of the perils as well as indicate some of the rewards that can be found in this new field. The volume should discourage vague and unsupported criticism, just as it should encourage more solid criticism by those who sympathize with the basic problems at hand (p. viii).

In the present writer's mind, calling in an "outside" critic to make an "appraisal" of the methodological status of polling is analogous to calling in an independent certified public accountant to make an appraisal of the financial status of an institution. If the CPA came up with a total of the debits alone, one would feel that he had misconceived his task. And if, moreover, the total of debits was in error, even this partial result would be of questionable value. It is the feeling of the writer that McNemar's critique, despite some insightful remarks and valuable suggestions, approaches the above situation rather too closely to be considered a really constructive criticism of the present methodological status of public opinion measurement.

BIBLIOGRAPHY

1. CANTRIL, H. *Gauging public opinion.* Princeton: Princeton Univ. Press, 1944.
2. DODD, S. C. On reliability in polling: a sociometric study of errors of polling in war zones. *Sociometry,* 1944, **7**, 265–282.
3. JENKINS, J. B. Dependability of psychological brand barometers I. The problem of reliability. *J. Appl. Psychol.*, 1938 **22**, 1–7.
4. KING, M. B., JR. Reliability of the idea centered question in interview schedules. *Am. Sociol. Rev.*, 1944, **9**, 57–64.
5. MCNEMAR, QUINN. Opinion-attitude methodology. *Psychol. Bull.*, 1946, **43**, 289–374.
6. NEWMAN, S. H., BOBBITT, J. M. & CAMERON, D. C. The reliability of the interview method in an officer candidate evaluation program. Amer. Psychol., 1946, **1**, 105.

SOME PRINCIPLES OF ATTITUDE-MEASUREMENT: A REPLY TO "OPINION-ATTITUDE METHODOLOGY"[1]

HERBERT S. CONRAD

College Entrance Examination Board, Princeton, N. J.

INTRODUCTION

In a recent review, McNemar (7) turns his critical talents to the methodological aspects of studies of opinions and attitudes. Since some work in which the present writer took part has been honored by McNemar's attention, a response to criticism seems in order. In this reply we shall be concerned more largely with matters of methodological principle than factual detail. The reply is, for the most part, restricted to McNemar's comments on three articles only (1, 2, 12). A reply to other portions of McNemar's review has already been made by Crespi (3); and doubtless further replies may be expected from other sources.

Of the three articles principally dealt with in the present "Reply," the first (12) makes use of Harding's (5) "Scale for Measuring Civilian Morale," and of a specially prepared Questionnaire; this Questionnaire includes a variety of items which—on one hypothesis or another—were considered as possibly related to the characteristics tapped by Harding's "Scale." The second and third articles (1, 2) present results from the application of two scales for the measurement of war-optimism. The first of these scales consists of 10 items designed to measure "optimism with regard to the military prospects for an early, easy victory"; the second consists of 14 items designed to measure "optimism concerning the consequences of the war." In addition, for comparative purposes, a 24-item scale "designed to measure general or personal optimism" was applied. Longer scales would very likely have been desirable, but could not be employed because of lack of time.[2] We shall consider, first, various points related to McNemar's discussion of the second and third articles.

[1] The present article is a partial reply to the critical review by McNemar (7), entitled "Opinion-Attitude Methodology." The writer is indebted to those who have read this reply (in whole or in part) prior to publication: viz., Harold O. Gulliksen, Hadley Cantril, Ledyard Tucker, Leo P. Crespi, Joseph M. Miller, Irving Robbins, Richard Pearson, and Donald A. Peterson. Responsibility for the statements and viewpoints in the present article rests, of course, solely with the writer.

[2] In addition to the scales on war-optimism and general or personal optimism, the Miller (8, 9) scale of 18 items was administered; the results from this scale, however, do not enter into the present discussion.

The Uni-Dimensional Scale

One of the points repeatedly stressed by McNemar is the desirability of a "uni-dimensional scale." In view of the importance of this concept it is unfortunate that McNemar fails to define "uni-dimensionality" explicitly in statistical terms. From a common-sense point of view, however, it seems clear that any attitude-scale which is "uni-dimensional" in a psychological sense would have to be composed of items which are homogeneous, in respect to the quality being measured. If the component items are not homogeneous, some items would be measuring qualities not measured by others, and additional dimensions would thus be introduced. A certain degree of heterogeneity among items is, however, tolerable; provided, first, that the heterogeneity is non-systematic —and thus tends to be self-counterbalancing; and provided, second, that the scale is sufficiently long to permit effective operation of the counterbalancing tendencies. The first proviso implies that the items measure *one common factor;* or, at least, that the items measure one common factor mainly—i.e., the saturation of each item with the common factor is comparatively high. Obviously, the higher the intercorrelation of each item with every other item in the scale, the higher the common-factor saturation of each item is likely to be, and the more perfectly homogeneous or "uni-dimensional" the scale.[3]

From the complex origins and complicated nature of many opinions and most attitudes, we should judge that strictly uni-dimensional scales in the realm of opinions and attitudes may be virtually impossible to construct—except possibly for issues which are indeed quite narrow and simple. We agree with McNemar that the uni-dimensional scale is both a legitimate and a desirable scientific goal; but it is not the *only* legitimate and desirable goal. If the ideal of the uni-dimensional scale restricted attitude- and opinion-research to only the narrowest and simplest of issues, then this ideal might have the unhappy distinction of achieving scientific rigor at the cost of social usefulness.

Perhaps the practical upper limit of homogeneity among attitude-

[3] Some statisticians would wish to correct the raw intercorrelations among items for response-errors and for errors-of-grouping. Response-errors can be handled by correction for attenuation; and grouping-errors, by formulas that correct for the influence of broad categories.

It may be worth noting at this point that, if the attitude scale includes group-factors, the effective counterbalancing of heterogeneity is impeded—except in the improbable event that the group-factors are numerous, small, and negatively intercorrelated. (A general negative intercorrelation among group-factors is improbable or impossible; because if factors *A* and *B* are both negatively correlated with factor *C*, then *A* and *B* will tend to be positively correlated with each other.)

or opinion-items is the degree of homogeneity found among items in intelligence tests. Let us assume that the split-half reliability (corrected by the Spearman-Brown formula) for an acceptable 100-item intelligence test is .90. Applying the Spearman-Brown formula "backward"[4] we estimate that the average intercorrelation among items is only .082. Does this seem fantastically low? Not when one considers that individual intelligence-test items ordinarily correlate with the *total* test only about .30–.50. Although some voices have been raised against the lack of homogeneity among intelligence-test items, we are not aware that any one has gone so far as to say, with regard to such tests, that "the meaning of total scores is enigmatic." Yet McNemar makes such a statement about attitude-scale scores, when the average intercorrelation among the attitude-items is as high as that among items of acceptable intelligence tests. To quote (7, p. 363):

> One other finding of Conrad and Sanford should be noted: they estimate[5] that the average intercorrelation among the military optimism items is .08, among the "consequences of the war" items, .12, and among the personal or general optimism items, .07. Evidently . . . the meaning of total scores on each of these aspects of morale is enigmatic.

We suggest that the "enigma" which confronts McNemar is largely an aprioristic one, springing from the application of standards which, if rigidly and inflexibly applied, would outlaw most intelligence-test scores, and ban nearly all (if not quite all) research on attitudes.

While the writer agrees with McNemar that uni-dimensional or very highly homogeneous scales are desirable (to the extent that they can be achieved), he definitely differs with the view that *only* strictly homogeneous elements may be combined into a single total score. To predict average academic achievement, for example, we need have no compunction about combining high-school grades, intelligence test scores, measures of lecture-note writing, and measures of academic drive—even though the correlations among some of these variables may be low or zero. To determine a man's income (whether financial or "psychic") we need have no compunction about adding whatever properly falls under income, whether the receipts from different sources are positively cor-

[4] In his review, McNemar appears to sanction this use of the Spearman-Brown formula. The formula is not strictly applicable in the present instance, because it assumes that all intercorrelations among items are equal, and that the standard deviations of responses to each item are alike. We judge that the present application of the Spearman-Brown formula results in a figure which is somewhat too high—but accurate enough for present purposes.

[5] This estimate was made by a method more accurate than the Spearman-Brown formula (see 1, p. 301).

related, negatively correlated, or totally uncorrelated. A salesman's financial gain, for instance, from one line of goods may be entirely uncorrelated with his gain from another: but the two types of gain belong together, because they relate to the question we are asking: viz., "How much money does he make as a salesman?" Similarly, responses (in terms of assent or dissent) to the item, *After this war there will be more hatred among peoples and nations than there ever was before* may or may not be highly correlated with responses to the item, *If a League of Nations is formed after this war, it will probably function with much the same inefficiency and impotence as the League of Nations after World War I:* but the items belong together, because they both relate to the question, "Is this person characterized by optimistic or pessimistic views regarding the consequences of the war?" Uncorrelated aptitude measures should be combined if they have the common characteristic of predicting the ability in the question; dollars from different sources (whether correlated or not) should be combined if the dollars have the common characteristic of adding to financial income or purchasing power; and responses to different items should be combined, if they have the common characteristic of pertinence to the personal trait or attitude under consideration. When McNemar (**7**, p. 298) writes that the "haphazard adding of dissimilar parts on the assumption that a meaningful whole will result is nonsense," he is, of course, right—but he has "stacked" his case by the terms "haphazard" and "dissimilar." We are referring to the well-considered addition of similar parts.

Every one will grant that the meaning of a score from a unidimensional or comparatively homogeneous scale is more clear-cut than that from a multi-dimensional. But if—as often seems the case—we are confronted with a multi-dimensional phenomenon, it may be excessively cumbrous to attempt to develop a reliable scale for each of the numerous dimensions. For one thing, the subjects would rebel: they cannot spend all their time being measured. In such a case, no matter what may be our theoretical approval of the uni-dimensional scale, the multi-dimensional scale—which is, however, uni-dimensional in the *a posteriori* sense that each item relates to the question at issue—becomes necessary.

If we read aright, McNemar appears to recognize this situation (temporarily) when he writes (**7**, p. 300):

> The requirement that a scale shall reflect a unitary attitude does not rule out the possibility of attempting to measure general attitudes or complex attitudes or specific attitudes. The complex ones would need to be broken down into component parts and a scale provided for each part. *It may be that some so-called attitudes will not be amenable to scaling as unitary characteristics.* (Italics ours.)

This quotation is from page 300. By the time McNemar has reached page 361, however, his tolerance for complex attitudes (or "so-called attitudes") seems to have vanished. He points out that a "life-satisfaction index" of Watson's is based on "markedly disparate" parts—and as such is "unequivocally meaningless." It seems to us that McNemar fails to distinguish between the "disparate" (the absolutely and essentially different, the incongruous or incompatible) and the merely heterogeneous. "Life-satisfaction" may well be a truly heterogeneous or multidimensional affair—and as such, require (or at least permit) the combination, into a single index, of satisfactions from widely varied and uncorrelated sources. Granted that any single "life-satisfaction index" suffers shortcomings, it may still, for certain purposes, be more useful than any other measure.

Let this be added. Items which are, from one point of view, combinable into a single scale, are not on that account barred from consideration as separate entities. Intelligence and academic drive are combinable into a single index for predicting school achievement. This does not imply that the distinction between intelligence and academic drive need be disregarded or lost. The fact that two aspects of academic ability may be combined is not a valid argument against studying the two aspects separately. Similarly, the fact that, say, military-optimism items are combinable to yield a military-optimism score does not argue against the desirability of studying the military-optimism items separately.[6] McNemar considers that such a two-fold approach "creates faith in neither the meaning of total scale scores nor what can be deduced from item analysis" (7, p. 360).

Reliability of Measurement

An early criticism of McNemar's (7, p. 362) is that the "form-vesus-form reliability coefficients" of the two war-optimism scales are "low": .45 for the 10-item scale, and .62 for the 14-item scale. (The corresponding split-half reliabilities, as McNemar mentions, are somewhat higher, .49 and .68.) Our reply is three-fold:

a. McNemar makes no contribution by pointing out what is already fully stated and elaborated in the original paper.[7]

b. The split-half reliabilities of the scales do not compare unfavorably

[6] The need for studying responses to individual items has been set forth by (among others) Miller (10), Jones (6), Dudycha (4), and Conrad and Sanford (2).

[7] From the original paper: "It is apparent that the reliability coefficients . . . are rather low. These reliability coefficients are not high enough to justify case-studies—except possibly for individuals making extremely high or low scores, or for cases on whom much supplementary information concerning war-attitudes is available. The reliabilities are, however, adequate for group studies cuch as the present" (1, p. 299).

with the split-half reliability of other similar scales such as Harding's (5) and Miller's (8, 9), when the number of items is taken into account (1, pp. 299–300).

c. Reliability coefficients cannot properly be judged "low" or "high" on any absolute basis. A test-reliability of .95 would be *low* if a man's life or death depended solely on his test score. By general agreement, a reliability of about .50 is adequate for the study of groups (the larger the group, the lower the permissible reliability). McNemar, to be sure, appears to reject "the 18-year old dictum of one of our better-known statisticians to the effect that a reliability coefficient of .50 is sufficient for group comparisons" (7, p. 295). The only reason offered for this view, however, is that "far too many investigators have hidden their heads in the quicksand of that dictum." We are less impressed by such rhetoric than we would be by specific, empirical data.

It may be of interest to observe that, if the 10-items of the military-optimism scale were expanded to 100 items (of similar merit and similar statistical characteristics as the original 10), the form-versus-form reliability would (according to the Spearman-Brown formula) become .89. This is as high as the form-versus-form reliability of acceptable intelligence tests. Similarly, if the 14 items of the war-consequences scale were expanded to 100, the reliability would become .92. These figures are not presented to persuade either ourselves or the reader that .45 = .89, or .62 = .92; but simply to point out that the internal consistency among items in the war-optimism scales is certainly not very different from that of items in acceptable intelligence tests.

The "Neutral Point"

One conclusion from the study of the war-optimism scales had to do with the lack of optimism regarding consequences of the war (the mean response was roughly .8σ below the neutral point) (1, p. 292). In connection with this conclusion, McNemar criticizes the "tendency to assign absolute meaning to the mid-scale point as indicating a neutral position" (7, p. 362). His thought is that "changing the wording or form of the items can easily lead to a shift of means." This criticism, in the present instance, does not seem to carry much weight. In the first place, notable shifts (either up or down) as a result of "changing the wording or form of the items" are much more likely to occur in the case of *single items*, than in the case of a mean based on 14 items. As McNemar himself has said, when attitude-scales (rather than single questions only) are used, "the whole worrisome problem of bias [due to question-form and phrasing] nearly vanishes" (7, p. 327). Second, in the absence of specific criticism, we see no need for changing either the wording or form of the items. Third, there can scarcely be any doubt that for the items, as phrased, the "mid-scale point" *does* indicate a neutral position. Consider, for example, the item:

BEFORE THE END OF THE WAR, IT IS LIKELY THAT OUR TROOPS WILL BE AFFLICTED BY SOME SERIOUS EPIDEMIC, SUCH AS THE INFLUENZA EPIDEMIC IN WORLD WAR I.

Strongly agree *Agree* *Undecided* *Disagree* *Strongly disagree*

It appears self-evident that agreement with this statement indicates a pessimistic view (with respect to this item); disagreement, an optimistic view; and "undecided" (the mid-scale point), a neutral view. Other items are similar, and a similar interpretation of the mid-scale point is appropriate.[8]

Dispersion, and the Comparison of Dispersions

In the first study of war-optimism (1), comparison was made between the standard deviations of responses to the war-optimism scales and the scale of general or personal optimism. (Due adjustment was made for the different number of items in the different scales.) It was found that individual differences in war-optimism were fully as great as in general or personal optimism concerning matters not particularly related to the war. This finding, in conjunction with the findings regarding central tendency, was considered to "suggest a type of morale in the present sample which is skeptical and individualistic, rather than enthusiastic, inspired, and unified." McNemar challenges this conclusion, on the ground that the units in the scales for war-optimism and general or personal optimism may not be comparable. Unfortunately, there is no generally known or established procedure for proving either comparability or incomparability of units in different scales. McNemar does not adduce any specific reasons for assuming serious dissimilarity of units in the present case. The following facts, though not conclusive, all tend to favor the likelihood of equal (or reasonably equal) units: the items in the different scales were all cast in much the same form; they required the same form of response; they were constructed (with few exceptions) by the same authors: and they were administered at the same time: Further, the standard deviations of responses to individual items in the three scales are similar.[9]

[8] McNemar's mention of the "form of the item" helps recall the fact that all statements in the war-optimism scales were phrased pessimistically (*dis*agreement with the statement indicating optimism). Rundquist (11) has noted a clear tendency for the mean scores on such statements to tend toward the favorable (optimistic) end of the scale. As mentioned in the original paper (1, p. 305) this renders the rather low level of war-optimism in the sample all the more significant.

[9] The mean of the standard deviations for items on military optimism is .99, on war-consequences optimism, 1.00, and on general or personal optimism, 1.01. The corresponding medians are .98, 1.01, and 1.00. [The figures for military optimism are taken from unpublished calculations; those for the other two scales are given in a footnote (1, p. 294).]

Additional evidence on lack of unanimity of attitudes was drawn from the distributions of responses to individual items (2, pp. 164–167). It was found that, "in general, there are significant frequencies over at least four points of the total five-point scale. The fact that nearly half the distributions are bimodal emphasizes the lack of agreement in the present sample" (2, p. 173). McNemar is especially critical concerning this evidence. He writes:

> In the article on individual items it is again concluded that unanimity of response is lacking as judged by the spread of responses over as many as four of the five possible points for an item. The variance of this spread is about 1.00; now if we accept their own estimate of item reliability (or unreliability) as being less than .25, it can be said that at least 75%, perhaps 80%, of the lack of unanimity is due to response or measurement errors. If we reversed the Brown-Spearman formula, we would find .078 as the item reliability for the military optimism items and .115 for the "consequences" items.[10] Maybe as much as 90% of the lack of unanimity is a function of response errors!

McNemar's criticism is, we think, seriously in error. The correction of the variance for unreliability of measurement is applicable only if we wish to know the variance of statistically "true" responses (a "true" response being defined as the average of an infinite number of responses by the individual, measured and re-measured under identical circumstances). For present purposes, however, what we want to know is each individual's *current* attitude—of which his response to the item seems the best measure obtainable. Current attitudes, thus measured, can be denied validity only if it be assumed that item-responses are indeed evanescent, changing in random fashion from hour to hour or one day to the next.[11] Although this is conceivable, it does not seem likely. Given competent measurement, the *short-term* re-test reliability of the individual item is almost undoubtedly high. Divisive propaganda, it must be noted, feeds on such divergences as it finds "here and now." From this point of view, a current division of opinions or attitudes regarding the war was significant.—This is not, of course, to condone gross errors of measurement (such as might be caused by poorly designed directions or careless interviewing), nor to accept without question differences of response caused merely by unclear, ambiguous items.

[10] The inference seems to be implied that the "item reliability" of .078 and of .115 was discovered by McNemar and overlooked by the authors. Actually, these figures are in the original paper; and as there indicated, the figures were obtained not by reversing the Spearman-Brown formula, but by a somewat more accurate method (1, p. 301).

[11] There are sometimes important *real* changes from one day to the next: thus, the attack on Pearl Harbor changed American attitudes toward the war immediately. But such real changes have nothing to do with the adventitious variation or "response error" to which McNemar refers.

In the absence of such sources of error, however, it seems to us that when one person agrees (for example) that "*the selfish devotion of people to their own 'national interests' will lead to fresh international conflicts and new wars*," while another person disagrees—we have a difference of opinion which is significant. The first person expresses a pessimistic or cynical view, consistent with a Nazi-like policy of unbridled nationalism and "power politics"; the other expresses an optimistic view which facilitates the relaxation of narrow nationalism and promotes the settlement of issues by justice instead of power "show-downs" between competing nations. It is hard to believe that these two individuals are, at the time of questioning, "truly" alike in their opinions on this particular item; nor—more generally—can we believe that a (frequently bimodal) distribution of responses over four or five points of a five-point scale really indicates virtual unanimity of current opinion about an item. We are inclined, rather, to agree with McNemar's statement at another point in his review, that "variation within groups indicates the relative homogeneity [or heterogeneity] of groups in their opinion about an issue" (7, p. 328).[12]

Part-Whole Correlation

A comparatively minor point relates to the statistical interpretation of a part-whole correlation. McNemar writes (7, p. 363):

> Some of the correlations reported by Conrad and Sanford are of interest. Military optimism and "consequences of the war" optimism correlate .33, or .57 corrected for attenuation, thus showing that the two scales have something in common. The correlation between military optimism and the combined scale (sum of the two) is .73; whereas "consequences" correlates .88 with the combined scales. The difference between these two coefficients was explained (1, p. 304) on the basis of the larger standard deviation for scores on "consequences." A logical and more pointed explanation is the difference in degree of spuriousness when scores on 10 items are correlated with scores based on these 10 plus 14 others, as compared with that when scores on 14 items are correlated with scores on these 14 plus 10 others.

The reply to this criticism is that McNemar is mistaken. If the set of 10 items had a higher standard deviation than the set of 14, then the

[12] Even if a correction of variance for unreliability of measurement were desirable, it would have to be based on the short-term re-test reliability of the individual item (rather than the correlation of the item with other items). By "short-term" re-test reliability we mean alternate-hour reliability, or, at most, alternate-day reliability. We do not know the magnitude of the short-term re-test reliability for any of the optimism-items, but (as indicated in the text) assume it to be high. Of course, there might be some differences in re-test reliability from one item to another; if so, any single correction or generalization (such as McNemar has suggested) would not be specifically applicable.

10-item set would have a higher correlation with the total of 24 items. This seems almost self-evident; it can be proved by the following simple derivation:

Let A = a component of T, the total score.
Let B = a second component of T, the total score.
$T = A + B$

Then

$$r_{at} = r_{a(a+b)} = \frac{\Sigma a(a+b)}{N\sigma_a\sigma_{a+b}} = \frac{\sigma_a + r_{ab}\sigma_b}{\sigma_{a+b}}$$

$$r_{bt} = r_{b(a+b)} = \frac{\Sigma b(a+b)}{N\sigma_b\sigma_{a+b}} = \frac{\sigma_b + r_{ab}\sigma_a}{\sigma_{a+b}}$$

and

$$r_{at} - r_{bt} = \frac{(\sigma_a - \sigma_b)(1 - r_{ab})}{\sqrt{\sigma_a^2 + \sigma_b^2 + 2r_{ab}\sigma_a\sigma_b}}.$$

As is obvious from the formula, the difference $r_{at} - r_{bt}$ is positive (i.e., r_{at} exceeds r_{bt}) whenever σ_a is greater than σ_b—unless, of course, $r_{ab} = 1.00$, in which case A and B would obviously correlate equally with *any* other variable. Thus the difference betwen r_{at} and r_{bt} is a function of σ_a and σ_b, and not of the number of items in A or B. Of course, σ_a and σ_b will ordinarily depend, in part, on the number of items comprising A and B, respectively. But there is no necessary relation between the number of items in a test and its standard deviation (this depends not only on the number of items, but also on the standard deviation of responses to the items, and the intercorrelations among the items). In the case under discussion, $r_{at} = .73$, and $r_{bt} = .88$—and the difference between these two r's did not seem likely accountable merely by the excess of 4 items in component B.

The Term, "Scale"

McNemar has criticized our use of the word "scales." He writes: "To call these collections of items 'scales' seems unwarranted, since nothing is said concerning how the items were selected or whether any were eliminated." Although this criticism has justification, we should like to urge that a "scale" is better judged by what it does than by how it was derived. If a collection of items succeeds in arranging individuals (or groups of individuals) in appropriate rank-order, the items may pass muster as at least one type of psychological "scale" (this is McNemar's own criterion) (7, p. 294). In further criticism, McNemar says "Apparently the old *a priori* method [of scale-construction] of the 1920's was used." We would remark only that McNemar's criticism is itself *a priori*. The question is whether the scales have as much reliability as might reasonably be expected from such small groups of items, and

whether they provide useful information. The answers to these questions are, we think, intimated or insinuated in McNemar's review, but given explicitly in the original papers (**1**, **2**).

Meaning of "Morale"

The last point in the discussion of McNemar's criticisms of the war-optimism papers relates to a matter of semantics. McNemar says (**7**, p. 363):

> A correlation of .60 is reported for the combined "war optimism" scale with D. C. Miller's national morale scale. If this were corrected for attenuation it would become about .85, thus demonstrating that two of the many morale scales so far discussed really tap the same thing. What that thing is, is not made clearer when one notes that an individual was considered to have low morale if he believed in 1942 that "Germany will probably bomb" our industrial centers in the East. Such an opinion would be tantamount to *high* morale in the sense of [denoting] confidence in leaders, since a desperate effort was being made to inculcate that very belief.[13]

The argument in the last two sentences hinges on the equating of war-optimism and morale. But the Conrad-Sanford scales do not claim to be "morale" scales. It was clearly recognized that "war-morale is much too complicated a phenomenon to be adequately diagnosed merely from the measurement of war-optimism" (**1**, p. 310). (McNemar has himself objected to "calling a wide variety of things by the name of morale") (**7**, p. 365). The comments, then, on "Germany will bomb" seem irrelevant. It cannot be denied that the belief that "Germany will bomb" was *pessimistic*.

Selection of Items and Conditional Interpretation

We turn now to a consideration of McNemar's criticism of the paper by Sanford and Conrad (**12**) which makes use of the Harding scale (**5**). McNemar writes (**7**, p. 361):

> In an effort to find some of the personal-social correlates of morale, Sanford and Conrad gave the Harding scale to 100 men and 173 women in a course in Mental Deficiency (what universe this sample represents is not specified). The report is based mainly on the results for men and presents the relationship of morale to 12 of 53 items in a questionnaire which the subjects also filled in. The authors do not reveal their reason for choosing to report on these particular [12] items . . .

[13] We question whether the war leaders aimed to inculcate a *belief* that "Germany will probably bomb"; they aimed, rather, to excite *fear* of the possibility that Germany *might* bomb. It was to forestall this possibility (among others) that maintenance and heightening of the war effort were demanded.

The last statement is simply not true. We quote from the article (12, p. 5):

> The questionnaire is presented in full below. Items marked by an *asterisk* are discussed in the present paper. Items marked by a *dagger* (†) were investigated to the extent of ascertaining the mean difference between groups; the differences found, however, were too small to justify discussion or elaboration. (The possibility, of course, remains that these items might prove significant if considered in relation to other information.) Items marked with a *small zero* have not yet been studied, either because these items did not appear promising, or because they did not lend themselves readily to quantitative analysis. It may be noticed that some of these "small zero" items are similar or related to others in the Questionnaire for which statistical analysis has been carried out.

We are at a loss to understand how McNemar could have overlooked this entire paragraph.

Regarding the 12 items mentioned above as selected for study, McNemar goes on to say: " . . . if these 12 items were the only ones which yielded relationships approaching statistical significance, we obviously have a hazardous capitalization on chance." That the authors were well aware of this point and had themselves pointed it out, is clear from the following statement in the article: "The facts summarized above represent, in general, a selection of the more interesting and promising of the available findings. Errors of sampling may have rendered some of the results more positive than they really are. On the other hand, added significance may perhaps be attached to such positive findings as were obtained, because of the limited reliability and the limited homogeneity of the Harding morale scale" (12, p. 20).

Theory-Formation and Significance-Tests

Like many others, McNemar does not consistently give due recognition to the evolutionary stages of theory-formation. In a young science, the first task is to "fish around" for hypotheses: we may call this the "hypothesis-finding" stage. During this stage, broad fact-finding (what McNemar calls the "drag-net" approach) is not merely appropriate, it is indispensable. One is fortunate if, at this stage, he can arrive at sensible hunches, and formulate them accurately and operationally. Later, one examines various facts (already available, or specially obtained) in order, first, to judge whether these facts conform with the tentative hypotheses previously developed; and, second, to form modified or new hypotheses. Finally—and typically much later—one may develop a crucial test (or a set of crucial tests—preferably experimental in nature) to determine, on a quantitative basis, the truth or limitations or falsity of a specific refined hypothesis. This last is the scientific "glamor-

stage" of theory-formation; and it is apparently to this stage that McNemar owes his chief allegiance and respect. Like many others, McNemar is interested mainly in the transformation of hypotheses into theory or law. In attitude-measurement, however, we are still at the first stage, or perhaps the second. In these early stages, the important point is *not to miss possibly good hypotheses.*

Evidently with thoughts like these in mind, Sanford and Conrad (12) prepared "complete scatter-diagrams, paying special attention to the extremes." (The lack of high reliability of the Harding scale (12, p. 17) suggests that only the extreme groups were adequately differentiated from each other.) It was noted that "although, for the cases around the middle, appreciable trends are not observable in the scatter-diagrams, interesting observations may be made with regard to differences between the extreme groups." It was found that—

> One of the most discriminating questions was. . . . "How many children would you like to have?" Here, as for many other items, there is of course considerable pressure toward a conventional answer: a man is expected to want to have *some* children. It is not very surprising, then, that the medians of the top- and bottom-scoring groups . . . are the same. Nevertheless, it is clear that the high-morale men tend definitely to differ from the low-morale group in their expressed desire for children. Three of the high-morale men want four or more children; only one[14] of the low-morale men wants this many. Two of the low-morale men want either one child or none; none of the high-morale men wants fewer than two children (12, p. 13).

The reader will recognize the "clinical," "hypothesis-hunting" nature of the remarks quoted above. McNemar is, however, sharply critical. He writes (7, p. 361):

> No significance tests are reported; our casual calculations indicate borderline or less than borderline (5% level of confidence) significance. For example, "one of the most discriminating questions" concerned the number of children desired by the subjects. Calculation shows that the correlation between this and scores on the Harding scale for morale is .06 for 88 cases. By no stretch of the imagination or by no amount of statistical juggling can so small a correlation be considered significantly different from zero; yet these authors state that "it is clear that the high-morale men tend definitely to differ from the low-morale group in their expressed desire for children." This inflation of a correlation of .06 is accomplished by considering only the 9 high and 8 low morale men and ignoring the question of statistical significance.

Since the authors had already indicated the absence of "appreciable trends" for the bulk of the cases, the coefficient computed by McNemar

[14] This individual was Italian, a good attending Catholic, and fundamentally if not frankly Fascist in outlook (13). All these factors are considered favorable to the expression of a desire for a large family.

may be regarded as confirmatory. It is true that figures on statistical significance were not reported, and that they could have been of service. It is also true, however—especially in the field of personality, where heterogeneous, multiple causation is typical—that some good hypotheses can easily be buried by the mechanical application of the null hypothesis and significance-tests. As McNemar himself has said: "When subgroups are being compared or when interrelationships are being determined, it is well to remember that the acceptance of the null hypothesis is particularly fallacious if the number of cases is small. Differences of practical importance can too easily be brushed aside because of lack of statistical significance . . . [When samples are small] the sampling errors become so large that it is indeed difficult to reject the null hypothesis even though sizable differences exist between the population parameters" (7, p. 337).[15] In brief, a small sample is not likely to give positive, unequivocal support to even the best of hypotheses. More than the immediate numerical facts enter into good early hypothesis-formation.

Choice of Sample

In his review, McNemar has disparaged the frequent use of college students as samples for the study of attitudes (7, pp. 331–333); in general, he prefers a cross-section of the total population (with some restriction as to chronological age). There is much to be said in favor of both types of sample. For our own part, we should consider an unselected (adult) sample of "mankind in general" as, usually, a poor sample for the scientific study of attitudes and opinions. In the first place, the use of such a sample limits, in practice, the number of questions that may be asked, and also restricts the nature of the questions—unless one is willing always to confine himself to "a cognitive level low enough for all respondents" (7, p. 318). Second, it is unlikely that a uni-dimensional scale would retain its uni-dimensionality when applied to a highly heterogeneous group: identical terms would not have uniform meaning for different parts of the sample, nor would identical responses have uniform significance. Third, few generalizations are likely to hold for all segments of a broad population—certainly, no generalization is likely to be equally applicable through all segments.

[15] The danger in this view, of course, is that an investigator may go "hog wild" in hypothesis-formation. We certainly hold no brief for small samples; neither, on the other hand, do we support the mechanical application and interpretation of significance-tests. In interpretation of the empirical facts and formal probabilities, judgment and insight provide both the brake and the spark.

Given a wide range of attitudes, social background, economic status, etc., relationships are quite likely to be curvilinear and heteroscedastic. Unless, then, an investigator has funds for a very large study—large enough to permit numerous subclassifications of the total sample, and to permit checks on the meaning of identical questions and responses in different sub-samples—he had better stick to one or two comparatively homogeneous groups. The very considerable advantages of college students as one of the comparatively homogeneous groups have been outlined in a previous publication (**2**, pp. 180–181); these advantages are by no means exclusively "practical" or budgetary. Of some pertinence to the papers on morale and war-optimism which McNemar has criticized is Miller's conclusion that "college students as a group are not distinguishable in national morale from the adults who live in the same areas" (**10**, p. 202).

Concluding Remarks

Although we have differed at several points with McNemar, we do not wish to be exclusively critical. McNemar has contributed a review which is distinguished for its alertness, its independence, and its unhackneyed character. Included in the review are several important constructive suggestions. In general, however, the tone of the review is definitely adverse: the general impression is one of widespread error and inadequacy. The tendency has been less to appreciate achievement than to criticize shortcomings. McNemar's evaluation fails to consider that many studies in the field of opinion-attitudes are basically oriented toward fact-collection: an issue is "hot" and some facts are wanted: only approximate accuracy is needed for the practical purpose, and the emphasis is on speed and economy: the scientific aspects and implications, frequently, are deliberately and necessarily subordinate. Whatever scientific contributions can be gleaned from such studies should be received cautiously, but not without gratitude.

What are the public-relations effects of a sharply critical review, by a distinguished psychologist, of an entire section of psychology? "Honesty is the best policy," but the present writer seriously questions whether McNemar's review has given a true picture of the practical usefulness and scientific status of the field of opinion-polling and attitude-measurement. Has not social science enough to contend with from biased critics outside its ranks, without supplying "free ammunition"? A public preoccupation with errors, shortcomings, and inadequacies can hardly be the most effective means for stimulating confidence and support. It would have been fairer and sounder to apportion attention to

both the merits and demerits of the reports which McNemar has discussed. For many, McNemar's review will appear as much an attack as an "appraisal" (7, p. 289).

The following quotation illustrates the unsympathetic "slanting" of McNemar's account (7, p. 335):

> A 1945 article by Cantril is of interest in connection with the problem of sampling. He compared the responses obtained by different polls when identical questions were asked at nearly the same time by two or more of the following agencies: AIPO [Gallup], OPOR [Princeton], NORC [Denver], and *Fortune*. The respective *N*'s were 3500, 1200, 2500, and 5000. The discrepancies between pairs of questions ranged from 0 to 12 [per cent], with an average of 3.24 for 99 comparisons. Cantril hails this as a "highly creditable performance" considering the "expected margin of error" of "3 or 4 per cent." One notes immediately from the given *N*'s that the discrepancies are larger than expected on the basis of mathematical error formula.[16]

It is obvious from the title of Cantril's article ("Do different polls get the same results?") and from the quotation itself, that Cantril's problem was whether the discrepancy between *different* polling agencies exceeded the normal margin of error for a *single* agency. One might have feared that different agencies would turn in markedly different percentages—considerably beyond the "expected margin of error." The findings justify Cantril's favorable verdict. The fact that the results for *each* agency have errors beyond the ideal, irreducible sampling-error is something else again. The polling agencies, in general, aim to achieve results which are *sufficiently* accurate for their purposes: and these purposes, in general, are not purely scientific. Cantril and McNemar have obviously addressed themselves to different questions. Both authors are right. But McNemar is wrong when he implies that Cantril has erred. To use a crude analogy, it is not wrong to say that alternate forms of an intelligence test (such as Forms *L* and *M* of the Revised Stanford-Binet) yield similar results, even though the discrepancies in IQ may be greater than accountable by pure theory. The unsympathetic "slanting" of the review is especially vexing, since McNemar is perfectly capable of well-balanced and helpful evaluation; thus, he writes about the Thurstone and Likert methods of attitude-scale construction

[16] While the standard error, as computed by statistical formula, is certainly a useful measure, it seems too much to hope that any practical, earthbound investigation will fail to include sampling errors over and beyond those envisaged by the formula. The question may thus be raised whether McNemar's use of the standard error as a criterion of sampling-adequacy is not excessively rigorous. Even sampling which (by all reasonable standards of excellence and permissible expense) appears quite careful and competent might yield fluctuations greater than those to be expected from the "mathematical error formula."

that "both methods have merits, and both have defects which might be overcome by a combination of the two" (7, p. 308). Statements of this character, unfortunately, are not typical of the review.

Some of the studies criticized by McNemar have, at least, the merit of emphasizing points which he has himself stressed. Thus, McNemar devotes considerable space to the limitations and dangers—from a scientific point of view—of the single-question poll. It happens that a paper in which the present writer had a part devotes an entire separate section to the "Inadequacy of the Single-Question Poll" (2, pp. 179–180). While this particular study was subjected to its fair share of adverse comment, nothing at all was said about this (at least *one*) commendable aspect. We believe that other authors could cite similar instances.

The study of opinions and attitudes will go on—more and more effectively, as continuing advances in knowledge and understanding are utilized. What is needed in this field, as in all others, is more time, more money, and more and better personnel—and judicious, well-balanced appraisal together with valid criticism.

Summary

The present paper is a reply to McNemar's (7) review entitled "Opinion-Attitude Measurement." The reply is concerned more largely with matters of methodological principle than factual detail, and is limited, for the most part, to McNemar's comments on three articles only (1, 2, 12). These articles deal with an application of the Harding "Scale for Measuring Civilian Morale" (12); a questionnaire prepared by Sanford and Ghiselli (12), containing items considered as possibly related to morale; the Conrad-Sanford scale for measuring "military optimism" (1, 2); and the Conrad-Sanford scale for measuring "optimism concerning the consequences of the war" (1, 2).

1. *The "uni-dimensional scale."* McNemar declares that the meaning of total scores on the two optimism-scales is "enigmatic," on the ground that the average intercorrelation among the items of each scale is low. The average intercorrelation among the optimism-items is, however, approximately the same as that among items of acceptable intelligence tests.

Although "uni-dimensional" scales are admittedy desirable (if they can be attained), the view that *only* strictly homogeneous elements may be combined into a single score is fallacious. Attitude-items with low intercorrelations are combinable into a total score if they have the common characteristic of pertinence to the personal trait or attitude under

consideration—just as aptitude-measures with low intercorrelations are combinable if they have the common characteristic of pertinence to the ability in question. The optimism-items fulfill the requirement of pertinence.

2. *Reliability of measurement.* McNemar has failed to support with evidence his rejection of "the 18-year old dictum" (7, p. 295) that a reliability coefficient of .50 is sufficient for group comparisons. The reliability coefficients of the two optimism scales are adequate for the study of groups, and for the study of contrasted individuals at the extremes of the distribution of optimism-scores.

3. *The "neutral point."* McNemar criticizes the "tendency to assign absolute meaning to the mid-scale point [of the attitude-items] as indicating a neutral position," on the ground that "changing the wording or form of the items can easily lead to a shift of means" (7, p. 362). This criticism, however, is much less likely to be valid for a total score based on a group of items, than it is for results from a single item.

4. *Dispersion, and the comparison of dispersions.* A study of the dispersion of responses to individual items is criticized by McNemar on the ground that raw standard deviations were employed instead of statistically "true" standard deviations. This criticism is in error, however, when what we want to know is each individual's *current* attitude—of which his response to the item is the best measure available. Current attitudes, thus measured, can be denied validity only if it be assumed that item-responses are indeed evanescent, changing in random fashion from hour to hour or one day to the next. The short-term re-test reliability of the optimism-items is not known, but is probably high. Current attitudes were important during the war, because divisive propaganda worked on what it found "here and now." Our purpose was to discover whether the sample's current attitudes were homogeneous and unified, or heterogeneous and divergent. For this purpose we are inclined to agree with McNemar's statement, at another point in his review, that "variation within groups indicates the relative homogeneity [or heterogeneity] of groups in their opinion about an issue" (7, p. 328).

McNemar challenges a comparison of dispersions on different optimism-scales because of the possibility of unequal units. The following facts, though not conclusive, all tend to favor the likelihood of at least reasonably equal units for the attitude-scales under discussion: the items in the different scales were all cast in much the same form; all the items required the same form of response; all the items (with few exceptions) were constructed by the same authors; the items in the differ-

ent scales were administered at the same time; and the standard deviations of responses to individual items in the various scales are similar.

5. *Part-whole correlation.* The correlation between part and whole depends explicitly on the standard deviations of the parts and the correlation between parts—and not (as McNemar has suggested) on the number of items in each part.

6. *The term "scale."* A "scale" is better judged by what it does than by how it was derived.

7. *Meaning of "morale."* McNemar's perplexity concerning the meaning of two attitude-scales appears to arise from a failure to distinguish between war-optimism and war-morale. "War-morale is much too complicated a phenomenon to be adequately diagnosed merely from the measurement of war-optimism" (**1**, p. 310).

8. *Selection of items and conditional interpretation.* McNemar's complaint that "the authors do not reveal" their reason for selecting certain items for statistical report is entirely without foundation. It is recognized, in the study under consideration (**12**), that sampling-errors might have lent spurious positiveness to the results; on the other hand, added significance could "perhaps be attached to such positive findings as were obtained, because of the limited reliability and limited homogeneity of the Harding morale scale" (**12**, p. 20).

9. *Theory-formation and significance-tests.* In the early stages of hypothesis-finding, the important point is *not to miss possibly good hypotheses*. Especially in the field of personality, where heterogeneous, multiple causation is typical, some good hypotheses can easily be buried by the mechanical application of the null hypothesis and significance tests. As McNemar himself has said: "[when samples are small] the sampling errors become so large that it is indeed difficult to reject the null hypothesis even though sizable differences exist between the population parameters" (**7**, p. 337). The danger in such a view, of course, is that an investigator may go "hog-wild" in hypothesis-formation. In interpretations of the empirical facts and formal probabilities, judgment and insight provide both the brake and the spark.

10. *Choice of sample.* Despite McNemar's espousal, a cross-section of the total (adult) population is usually a poor sample for the scientific study of attitudes and opinions. First, the use of such a sample limits in practice, the number of questions that may be asked, and also restricts the nature of the questions. Second, it is unlikely that a "uni-dimensional" scale would retain its uni-dimensionality when applied to a highly heterogeneous group. Third, few generalizations are likely to hold for all segments of a broad population. Of some pertinence to the

papers on morale and war-optimism which McNemar has criticized is Miller's conclusion that "college students as a group are not distinguishable in national morale from the adults who live in the same areas" (**10**, p. 202).

11. *Concluding remarks.* McNemar's review is distinguished for its alertness, its independence, and its unhackneyed character. Included in the review are several constructive suggestions. In general, however, the tone of the review is definitely adverse. McNemar's evaluation fails to consider that many investigations in the field of opinion-attitudes must place emphasis on speed and economy, and are basically oriented toward practical fact-collection rather than scientific advance. Scientific contributions from such studies should be received cautiously and critically, yet not without gratitude.

The public-relations effect of an adverse review, by a distinguished psychologist, of an entire section of psychology can scarcely be favorable. It would have been sounder to apportion attention to both the merits and demerits of the studies considered, and to frame criticism in the perspective of progress.

BIBLIOGRAPHY

1. CONRAD, H. S. & SANFORD, R. N. Scales for the measurement of war-optimism: I. Military optimism; II. Optimism on consequences of the war. *J. Psychol.*, 1943, **16**, 285–311.
2. CONRAD, H. S. & SANFORD R. N. Some specific war-attitudes of college students. *J. Psychol.*, 1944, **17**, 153–186.
3. CRESPI, L. "Opinion-attitude methodology" and the polls: a rejoinder. *Psychol. Bull.*, 1946, **43**, 562–569.
4. DUDYCHA, G. J. A critical examination of the measurement of attitude toward war. *J. soc. Psychol.*, 1943, **18**, 383–392.
5. HARDING, J. A scale for measuring civilian morale. *J. Psychol.*, 1941, **12**, 101–110.
6. JONES, V. The nature of changes in attitudes of college students toward war over an eleven-year period. *J. educ. Psychol.*, 1942, **33**, 481–494.
7. MCNEMAR, Q. Opinion-attitude methodology. *Psychol. Bull.*, 1946, **43**, 289–374.
8. MILLER, D. C. The measurement of national morale. *Amer. sociol. Rev.*, 1941, **6**, 487–498.
9. MILLER, D. C. How's your morale? *Amer. Mag.*, 1942, June, **133**, 91.
10. MILLER, D. C. National morale of American college students in 1941. *Amer. sociol. Rev.*, 1942, **7**, 194–213.
11. RUNDQUIST, E. A. Form of statement in personality measurement. *J. educ. Psychol.*, 1940, **31**, 135–147.
12. SANFORD, R. N. & CONRAD, H. S. Some personality correlates of morale. *J. abnorm. & soc. Psychol.*, 1943, **38**, 3–20.
13. SANFORD, R. N. & CONRAD, H. S. High and low morale as exemplified in two cases. *Charac. & Personality*, 1944, **12**, 207–227.

SEWARD, GEORGENE H. *Sex and the social order.* New York: McGraw-Hill, 1946. Pp. xii+301.

The increasing scientific interest in problems of sex has produced many investigations by specialists who have frequently remained more or less isolated from each other because of indifference or the complexity of and difficulty of access to published and unpublished knowledge in unfamiliar disciplines. A definite need has developed for a careful, evaluative survey of the whole gamut of sex studies. Dr. Seward has applied herself to this task with remarkable industry, with conviction that great possibilities for human welfare lie in a better understanding of sex, and with confidence that a social psychologist, with first-hand knowledge of research methods in comparative psychology, is in a most advantageous position for making the survey. The result, this solid volume of 301 pages and 701 references, is very satisfactory.

Approximately 35% of the references are from the psychological literature (half of these are animal studies), 25% from the biological, 25% from the psychiatric and medical, and 15% from the sociological, anthropological, and educational literature. The excellent organization of this vast material is based on a combination of the phylogenetic and cross-cultural approaches. Sex is given a broad and dual meaning: (1) mating behavior, in its sexual, reproductive, conjugal, and parental aspects; and (2) the social roles assigned to men and women. It is the hope of the author to achieve objectivity in examining our own culture through knowledge of the variations in sexual behavior and sexual role in animal and primitive human groups and to integrate the biological with the cultural data. A competent objectivity is achieved and is only occasionally impaired by the well-directed zeal of the author for social change. But the attempted integration of data is very superficial, as should be expected from the fact that we have as yet not a single study of sex, let alone survey, which brings together in a significant way the relevant biological, psychological, and social factors. Dr. Seward's brave attempt should be repeated when more extensive studies of sex have been made and reported.

The first chapter, entitled *Orientation*, is, surprisingly, by far the weakest. It refers to the evidence that control of sex is found in all cultures and suggests that this control sooner or later conflicts with certain physiological limits and that cultures may be rated with respect to frequency of behavioral disorder resulting from such conflict. Very inadequate reference is then made to a few theoretical and clinical approaches to sexual problems. A gratuitous promise to evaluate therapeutic methods is almost completely neglected thereafter.

The next six chapters constitute the best existing general review of

animal sexual behavior. The major topics are the dynamics of sexual behavior (hormonal and neural factors), the development of sexual behavior in relation to social organization, sexual behavior in rodents, maternal care in lower mammals, and conjugal relations and the family in primates. The author finds "clear and consistent evidence . . . showing dominance to be correlated with male sex hormones" and to result in a "natural line of cleavage" between sexes, which is complicated in the higher infrahuman primates by "social personality factors." The author also emphasizes (much too strongly, in the face of the evidence) the theory that sexual behavior becomes increasingly subject to social control as the phylogenetic scale is ascended. Careful definition and analysis of social control are sorely needed here but are not offered. An unwarrantable reference to a "cultural superstructure" in animals is made probably to strengthen the attempted integration of biological and social data.

Chapters 8 and 9 present a cross-cultural survey of (1) the function of sexual activity and (2) the sex typing of social role in at least 15 different nonliterate groups. "Implications for Western culture" are drawn from the variety in patterns of sexual behavior and in sex temperament: notably, there is discussion of "the false assumption that a dichotomy in sex temperament is somehow essential and desirable."

The next chapter is a brief history of changes in the status of women in Western culture. Prolonged wars are found to have permitted social, economic, and political gains for women, which have usually been lost after the men came home. The steepest male-female gradient is found in contemporary authoritarian groups, particularly in the Third Reich, where women were subjected to "a baser degradation than they had ever experienced in the history of Western civilization." In Russia, by contrast, socialism has produced greater sexual equality and has granted special compensations to married female workers. The author is cognizant of reports of recent reaction in Russia, but concludes: "Unless the economic structure reverts to capitalism, we need not fear a lowering in women's status."

Chapters 11–14, on human sexual behavior in childhood, adolescence, adulthood, and senescence, are concise, balanced, informative, and objective and constitute what is perhaps the most important section of the book. The author should be commended for the persistent search for important psychological factors in all of the problems discussed. The value of these chapters permits detailed mention of limitations.

Except in the discussion of those problems in which she has had experimental and clinical experience, notably menstrual cycle and menopause, the author is unable to present the second-hand quantitative data and incomplete clinical reports in a way which makes for better understanding of just how sexual behavior is related to the total personality of an individual. But it is unfair to expect the author to im-

prove on the material at hand, for it has been the unfortunate tradition in sex studies to take activity in just one or two sexual outlets, such as masturbation, nonmarital intercourse, or homosexual activity, or one sexual aberration and show its relation to personality adjustment. This tradition may not be corrected until many of the reports of the comprehensive Kinsey project on human sexual behavior have been published in the years to come. Obviously, the balanced pattern of the various sexual outlets (masturbation, dreaming to climax, petting to climax, marital and nonmarital intercourse, homosexual contacts, and contacts with animals), the change in pattern in the life-history, and the relation of the pattern to the special customs and sanctions of one's own group are more important than the occurrence or non-occurrence or the rate of occurrence of activity in a particular outlet which is arbitrarily isolated for clinical or research examination.

Criticism can also be directed

1. at the data chosen to show the prevalence of impotency in men (other data show it to be much less frequent than frigidity in women),
2. at the author's reference to the relation of the so-called androgen-estrogen balance to male homosexual activity (what is the functional significance of relative amounts of different hormones found in the urine of males, and what groups, if any, do the selected patients represent?), and
3. at the statement, suggesting a trend toward greater incidence of premarital intercourse, that "figures for premarital intercourse have risen from a low of 7% with Davis' rather conservative group in the 1920's to the 25% report by Landis . . . in the 1930's," (in the 1920's, Hamilton's excellent study found 35% incidence for women).

One misleading textual error was found on page 202. The Terman and Hamilton statistics for the indicated groups show coital frequencies of once or twice per *week*, not per *month*.

Whereas the recent Scheinfeld volume, *Women and Men*, emphasizes biological sex differences and their derivatives, the Seward chapter on this topic aptly emphasizes the socially imposed differences. The author suggests that the "natural line of cleavage" between sexes, related to strength and to social dominance in both primitive and slum conditions, has been encouraged and unnecessarily exaggerated by most human cultures. She concludes "that the scientific evidence available at the present writing fails to indicate differences attributable to sex membership as such that would justify casting men and women in different social roles." The final chapter, *Sex in Postwar Society*, develops this conclusion into a plan for a democratic "redefinition of the roles of men and women," involving social changes which would permit

1. women to learn to achieve and to grasp the same social opportunities given men,
2. men to learn to cherish and to participate in "life-giving" functions,
3. equitable sharing of domestic burdens,

4. education of children for social sex roles, and
5. elevation of feminine values.

The book is a definite contribution for those who are engaged in research on sex, who teach or study in courses involving sexual development, and for those who are consulted about sex problems. The incomplete and specialized nature of the scattered research publications on sex has led the author to use too many metaphors, analogies, and tentative statements of cause-effect relationships and a style suitable to an elementary text book. The chief fault of the book is failure to point clearly and frankly to the inadequacies and gaps in present knowledge; its chief virtue, the demonstration that society cannot afford to neglect further knowledge.

VINCENT NOWLIS.

Indiana University.

SCHEINFELD, AMRAM. *Women and men.* New York: Harcourt, Brace, 1944. Pp. xx+453.

If this popular book served no other purpose than its stated objective of helping "women and men toward a better understanding of themselves in relation to each other," it would well be worth reading. To the social scientist, however, it offers a more tangible service, for it brings together a large body of recent research findings on sex differences from the fields of anthropology, biology, medicine, psychology, physiology, biochemistry, genetics, criminology, vital statistics, and sociology. The facts presented are well documented, and each chapter has been critically read by one or more recognized experts in the areas concerned. The style is smooth, sprightly, non-technical, and probably within the range of comprehension of the average high-school student. Important points are often clarified and emphasized through use of tables and clever, often amusing, drawings and diagrams. Content is broad in scope. Early chapters are devoted to sex determination, prenatal developmental sex differences, and the sex-ratio. There follows a genetic treatment of sex differences from birth to adulthood in the fields of physical growth, morbidity and mortality, motor development, intelligence, social relationships, and personality. Other chapters deal specifically with sex differences in adults in sex life, crime, clothing, occupations, and creativity and genius. Social implications are discussed extensively in the concluding chapters entitled *Sex Equality*, *The Soviet Experiment*, and *Marriage of Tomorrow*.

Writing in a controversial field notable for the extremeness of the views held by various authorities, Scheinfeld seems to this reviewer to have maintained remarkable objectivity. Especially does he seem impartial in his treatment of the relative abilities and accomplishments of women and men, and in his handling of their problems of adjusting to the roles assigned to them by virtue of their sex. Essentially, the thesis

elaborated is that basic biological differences present from early prenatal life, and increasingly apparent as development proceeds, are predominantly responsible for behavioral differences in the sexes and for the differential demands made upon them by society. Cultural factors are consistently considered, but where they are operative, they are interpreted by the author as serving primarily to reinforce the native differences. To support his views he frequently resorts to analogy, citing experimental findings and observational reports on infra-human forms.

No doubt some readers will feel that Scheinfeld has overplayed the native influences. In this connection it is interesting to note the author's explanation of how he came to write the present volume:

> According to the original outline, I had expected to devote myself mainly to the social factors, past and present, as they have served to influence the relationships between the sexes, and to give only passing attention to biological sex differences. In this approach I was reflecting the prevailing tendency among social scientists to regard differences between women and men in behavior, thought, temperament, and achievement, as chiefly the products of "conditioning."
>
> But as intensive research proceeded, as pertinent facts were brought together and new avenues explored it began to appear that the original premise had many weaknesses. The basic sex differences, I was forced to conclude, were far more extensive, and had far more to do with the behavior patterns, capacities, and activities of the sexes, than most persons in professional circles had suspected or conceded. Further, in the light of these facts various widely current theories began to appear highly questionable (p. ix).

The principle criticism of *Women and Men* from this reviewer's point of view, is the author's emphasis on differences between the sexes to the exclusion of the voluminous research that has revealed no differences or only negligible differences. Perhaps this is to be expected in a book designed to reveal sex differences, but it is felt that the author might have devoted one chapter, at least, to pointing out the many behavioral similarities of the sexes. The book might also have been more useful to scientific readers had Scheinfeld presented studies in greater detail and with a statement of the statistical significance of the reported differences. In connection with this latter criticism, however, it seems only fair to point out that the volume was intended for the layman, not the scientist.

Considered in the whole, *Women and Men* is an outstanding contribution to the literature on sex differences. It can well be recommended to professional and non-professional readers alike; to all, in fact, who work with people or are interested in them. It does not seem amiss to add, further, that the level of scientific writing would be raised considerably if all authors took as extensive precautions as Scheinfeld has to assure the accuracy of the factual material presented.

MARGARET KUENNE.

University of Wisconsin.

GOODENOUGH, FLORENCE L., *Developmental psychology, an introduction to the study of human behavior* (2nd Ed. Rev.). New York: D. Appleton-Century, 1945. Pp. xxii+723.

A thorough revision of the first edition (1934) containing one hundred more pages, and, because of the larger size page, having approximately 40% more printed materials with many figures or graphs from the more recent literature. The format is decidedly more pleasing than that of the first edition, and in spite of the war-time restrictions on paper the print is more easily read. This edition contains a good analytical table of contents which is helpful to the reader. The organization of materials in the first edition was good but it is distinctly better in the second edition. The 1934 edition comprised 26 chapters; the 1945 edition, 30, many of them new chapters. The materials are grouped into five parts. Part I, *Principles and Methods of Modern Psychology*, includes two introductory chapters. Part II, *The Child's Equipment for Living*, contains five chapters presenting important developmental materials on hereditary background, prenatal development, the child at birth, and the arousing and patterning of activity. Part III on *The Normal Course of Human Development* constitutes the major portion of the volume. It includes 16 chapters on the period before speech, early childhood, middle childhood, adolescence, the college years, maturity, the individual at work, old age, and learning and retention. Part IV on *Personality Deviations* includes four chapters on common handicaps of normal people, mental disease, mental deficiency, and juvenile delinquency and adult crime. Part V on *The Mental Hygiene of Development* contains the three final chapters, one on bringing up children, one on increasing human happiness and efficiency, and a concluding chapter.

The book is intended for use as a basic text for the first course in psychology. If used as basic text in any subsequent course, there would be considerable duplication of materials.

As would be expected, the author firmly believes in the developmental approach for the first course, but she is not alone in this view as may be seen from a hasty survey of the tables of contents of recent texts in general psychology. On the value of the developmental approach in the first course (the only one taken by large numbers of undergraduates) the author says in the Preface.

> For many years I have felt that no sound understanding of human or animal behavior can be had without reference to its beginnings, its course of development, and the factors by which that course is influenced. Man does not come into the world full grown. His talents do not originate in the psychological laboratory, even though they may there be first reduced to quotients and deciles. Behavior is quite as much a matter of growth as is stature. Its qualitative variants and their permutations and combinations are beyond human reckoning; yet their organization and patterning are at all times unitary and

coherent. But because of these multiple aspects, it is easy to lose sight of the individual in our concern with his reactions. We reify man's "traits," and in so doing we forget the man. . . . Because the ghost of faculty psychology is hard to exorcise, it becomes essential for students to realize from the start that the fundamental behaviorial unit is not a depersonalized trait but a living individual (viii).

To the reviewer it seems clear that students who master this book will have a good foundation in the topics usually included in the first course since most of them are well-covered in it. In addition, however, they will have a better understanding and appreciation of human behavior just because they know the fundamentals of its development.

Each chapter is introduced by several italicized questions which direct the student's attention to important topics in it. They are pointed enough to be of distinct help to him in reading it.

At the end of each chapter is a one-page description of a book or class experiment throwing additional light on some important topic of the chapter—this in lieu of the list of references often found at the end of the chapters in undergraduate texts. Such lists indicate nothing of the author's scholarship and have very little value for undergraduate students.

The chapters are well-written with ample use of recent research findings. The book is well-adapted for use as textbook (or collateral reading) in the first course in general psychology.

FOWLER D. BROOKS.

DePauw University.

ALEXANDER, F. & FRENCH, T. M. *Psychoanalytic therapy. Principles and application.* New York: Ronald Press, 1946. Pp. xiii+353.

Diagnostic techniques in psychology have long left their mysterious grounds of intuition and finally moved into the safer shelters of the experimental laboratory, while the therapeutic methods have remained relatively static and untouched by the modelling chisel of controlled scientific experimentation. The present volume represents the results of seven years' experimentation in the field of psychotherapy. F. Alexander and T. M. French, together with nine staff members of the Chicago *Institute for Psychoanalysis*, were the experimenters. The impressive number of 292 cases treated at the Institute and an almost equally large number of patients seen in private practice bear witness to the fact that the findings are not based on traditionally offered isolated observations on a meager number of subjects.

Having defined mental disturbance as "a failure of the ego in performing its function of securing adequate gratification for subjective needs under the existing external conditions" (viii), the authors proceed to show the development of therapeutic techniques suitable to reestab-

lish the functions of a weakened or incapacitated ego. The shortcomings of the traditional psychoanalytic technique are pointed out in detail and the necessity for developing diagnostic methods that are economical both in terms of time and effort is stressed. This necessity for shorter techniques is dictated by the need for a psychotherapy which will not only be able to handle an ever-increasing number of emotionally unstable, but which also must encompass the vast number of mildly maladjusted individuals so abundantly present in our society. The validity of the traditional psychoanalytic method is not denied, neither are Freud's basic discoveries rejected, however, the authors feel that this technique is not only unsuitable for all cases, but may even prove dangerous and detrimental with certain types of patients. It is pointed out that the standard psychoanalytic method, as a therapy, tends to neglect individual differences and idiosyncratic needs of the patients. Some individuals may, in the long process of analysis, never learn to solve their problems, but rather develop a procrastinating kind of overdependency upon the analyst or substitute an intense preoccupation with past traumata for learning to realistically face their present pressing problems. Furthermore, it has been realized that the therapeutic process is not at all restricted to the "interview session" or the confines of the therapist's office. True, the essence of a successful treatment may hinge on a specific transference relationship between client and therapist, however, the patient has also other emotional relationships, with his wife, his family, his friends, and business associates. Thus, it has become the therapist's duty to *actively* utilize these other relationships, for "real life" situations may, at times, equal the interview sessions in therapeutic importance. Experimentation with the frequency of interviews and controlled temporary interruptions of the treatment has shown the tremendous therapeutic effect these outside relationships can have. The authors think of neurosis in terms of an "interrupted learning process" (74) at which the patient has reacted to problems in the past with a stereotyped behavior pattern which never led to a successful solution. Therapy, then, consists essentially of an "emotional reeducation" (95) where the patient learns to substitute new behavior patterns leading to a successful solution of old problems for old, unsuccessful attempts which have led to the neurotic symptoms. But more than that. Having learned these new behavior patterns, he must now be given an opportunity to integrate them by constantly applying them to "real life" situations. The proposed short psychotherapies in which the intensity of the transference relationship is actively controlled by the therapist and in which, through interruptions of treatment, the patient is forced to face his present, "real" problems, seem to fulfill these requirements.

The book is written in a fluid and interesting manner with an abundant presentation of case histories. Although the volume is divided into

smaller sections, the reader will be struck by the degree of integration that has been achieved. Occasionally, one may be bothered by some repetitiousness, however, considering the number of collaborators, this is perhaps unavoidable.

This work has shown, in a way, that psychoanalysis has not remained stale. New realizations and experimentation have caused it to grow and undergo modifications. Perhaps the most unique discovery of the present investigators was the fact that no single technique can dogmatically be used in all cases. A patient's problems and behavior peculiarities will not yield to any rigid, iron rules of a given treatment, but rather the nature of the therapy has to be tailored around the form of the particular problem. The "principle of flexibility" is the main theme of the presented therapeutic approaches.

The general psychiatrist may reject these short therapies as being still too analytically colored, while the psychoanalyst will probably accuse the authors of being too eclectic or even having left the psychoanalytic camp entirely. It must be remembered, however, that any therapy demonstrates its value in its practical usefulness and scientific soundness. Specific "party allegiance" must be subjected to academic and secondary interest.

LUDWIG IMMERGLUCK.

State University of Iowa
and Iowa Psychopathic Hospital.

BROCK, S. *The basis of clinical neurology* (2nd Ed.). Baltimore: Williams & Wilkins, 1945. Pp. xii+393.

Clinical neurology is, or should be, highly dependent upon the most recent research in neurophysiology and neuroanatomy in order that it may refine its diagnostic techniques and develop new therapeutic methods. During recent years there has been increased emphasis upon the functional, rather than the structural, aspects of neuroanatomical research and the advances in electrical recording and stimulating methods have given impetus to electrophysiological research. The result of these research trends, separately and conjointly, has been to increase greatly our knowledge of neural functions and integration. Recognizing this fact the author set himself the task of presenting "neuroanatomy and especially neurophysiology from the standpoint of clinical usefulness."

Although the author has succeeded in introducing some of the newer evidence from neurophysiological and neuroanatomical study the success of the approach attained in this volume must be judged marginal. Some of the shortcomings of this book in relation to its stated objectives are:

1. Neurophysiological, neuroanatomical and clinical data are set side by side with a minimum of interrelationship pointed out by the author.

2. Critical selection of materials and critical comment are notably lacking with the result that in several instances dubious concepts are presented and presumably espoused.

3. Careful reading reveals occasional errors, usually of omission or incompleteness of statement, but a more disturbing feature is the fairly frequent occurrence of loose and often meaningless statements.

The first two sentences of the introductory chapter are characteristic: "The study of mammalian and human neurophysiology has proceeded along various lines. The defect in function resulting from disease or injury has thrown much light on various problems."

The author is to be commended upon his organization and condensation of so much material and for the liberal use of good illustrations. Apt use has been made of tables and charts for condensing information on diagnostic signs, differential diagnoses, etc. Greatest emphasis has been given to neuroanatomy and clinical neurology, with a minimum of neurophysiology injected here and there. References are good, but not extensive.

The sections on the spinal cord and brain stem are especially good, as are also those on the basal ganglia and the vegetative nervous system. The treatment of the brain and its functions is spotty and little of the recent work on the thalamus and hypothalamus is given. A brief section on the electroencephalogram by Dr. P. A. F. Hoefer is included in Chapter 17.

The book should prove useful, especially for review purposes, if there is already familiarity with basic principles of neurology and physiology. It seems doubtful that it is adapted to the needs of the beginning student in neurology as the inside leaf of the paper cover intimates.

DONALD B. LINDSLEY.

Northwestern University.

AMERICAN PSYCHOPATHOLOGICAL ASSOCIATION. *Trends of mental disease.* New York: King's Crown Press, 1945. Pp. 114.

This series of six research articles by different writers constitutes an important addition to statistical knowledge concerning mental disease. The first four articles consider past, present and future trends with respect to the incidence of hospitalized mental patients in the general population. These studies are in general agreement that there has been and will likely continue to be a gradual but continuous increase in number of hospitalized patients, due mainly to a sharp increase in the incidence of old age psychoses. The latter increase is attributed to the increased longevity of the general population which has resulted in a larger number of aged individuals being exposed to the risk of old age psychoses. This, however, is an inadequate explanation. The figures are based on incidence rates per 100,000 population, which means that

a higher percentage of old individuals than formerly are being sent to mental hospitals. One possible interpretation is that individuals who are now living on to old age are more susceptible to mental disorders than in former years. A second and more probable explanation is that, in the past, families tended to take care of their aged relatives with mental symptoms at home, but at present, the trend is to send them to mental hospitals.

A more specific criticism is Malzberg's contention that during World War I there was a sharp increase in rate of first admissions to mental hospitals. Other investigators working with the same data have consistently failed to confirm this observation. The two remaining articles deal with World War II and mental diseases. The most significant point made is that most men who were rejected or discharged from military service because of psychiatric defects are capable of adequate adjustment in civilian life.

JAMES D. PAGE.

Temple University.

WEST, JAMES. *Plainville, U. S. A.* New York: Columbia Univ. Press, 1945. Pp. xviii+238.

This is an intimate study of a small rural community in the central part of the United States. The author, who writes under a pseudonym in order to insure the anonymity of the little village (population 275) in which he lived for a period of more than a year has dedicated the volume to Professor Ralph Linton, under whose direction the project was carried out. Although the point of view is primarily that of a sociologist, the report nevertheless has a good deal to offer to psychologists who are interested in community life and habits and in the impact of social change upon individuals whose former pattern of life had been relatively unaffected by outside influences.

The first chapter entitled, *Plainville*, gives the general setting. The little village of approximately sixty-five houses and a dozen small stores near the geographical center of "Woodland County" is located on a dirt road at some distance from a highway. The surrounding countryside is thinly populated and most of the inhabitants are acquainted with each other. Houses are shabby and the average income level is small. Subsistence is chiefly from farm products but no one lacks food. Bathrooms and central heating are practically non-existent but almost everyone has a car of some kind and a radio.

The two following chapters are entitled respectively, *Social Structure: General*, and *Social Structure: The Class System*. It is shown that in spite of the fact that the people place great stress upon the absence of class distinctions within the community, claiming that "this is *one* place where one man's as good as another," nevertheless there exists a caste

system that is, if anything, more rigid and inflexible than is likely to be found among social groups having a much greater range of income and education. Dozens of examples of the way in which this class system operates are cited, and the various practices by which children are indoctrinated with class consciousness and taught conformity to the social taboos are described in detail.

In the two following chapters, *Religion* and *From Cradle to Grave*, a detailed picture of the more intimate and personal aspects of Plainville life is presented. Of particular psychological interest is the account of typical attitudes displayed toward sexual matters and of the means by which children acquire their knowledge of sex.

The final chapter, *Plainville and the Future*, presents a brief overview of the culturally induced changes that have occurred in Plainville life since 1930 when the first settlers arrived and more particularly of those ensuant upon the coming of the many governmental agencies with alphabetical designations during the past fifteen years. The shock to traditional attitudes of "money without work" and of benefits conferred upon those who seemed least deserving is vividly portrayed. On the other hand it is shown that in spite of administrative bungling, the agricultural reform movement "may well prove to be the third in the series of great trait-complexes which, introduced from outside, have revolutionized the local society." And this is of far-reaching importance for "since there are millions of Plainvillers in America, the problem of Plainville is the problem of America."

FLORENCE L. GOODENOUGH.

University of Minnesota.

SIMMONS, LEO W. *The role of the aged in primitive society.* New Haven: Yale Univ. Press, 1945. Pp. vi+317.

There has long been a paucity of quantitative fact, sociological and psychological, on the status of the aged in various societal conditions. The present study therefore helps to meet a very real need. Data were selected from historical and anthropological reports (three hundred and sixty-six titles) concerning the treatment of the aged in primitive tribes. Seventy-one tribes in all were included, distributed throughout all parts of the world, most of them in the torrid and temperate zones; and one hundred nine societal characteristics or traits were chosen for quantitative treatment. The traits are classified under three heads: (1) Habitat, maintenance, and economic status; (2) Political and social organization; (3) Religious and miscellaneous beliefs and practices.

Each of the seventy-one tribes was rated on a four point scale with respect to each of the cultural traits, and co-efficients of association were computed by the Yule method. Over eleven hundred co-efficients were thus obtained.

There are eight chapters besides the appendix. They discuss the position of the aged with respect to food, property rights, prestige, general activities (economic functions and personal services), political and civic activities, the use of knowledge, magic, religion, the functions of the family and reaction to death. Most of this material is in the form of references to the anthropological data, with liberal quotations from sources. One is impressed by the vast amount of information which the author has condensed into his treatments.

The social psychologist will find the chapters on prestige, political and civil activities, and the functions of the family particularly interesting and informative.

One defect of the work as the reviewer sees it is that at certain points the case material and the quantitative data have not been adequately integrated. Thus in the chapter on prestige there are twenty-nine pages of case material, and then at the end there are two pages which discuss the co-efficients of association but without any clear attempt to relate them to the case material. On the other hand, in certain of the chapters a very effective use has been made of the quantitative data, and this is particularly true in the chapter on political and civil activities and the chapter on the functions of the family.

Other limitations are the lack of preciseness in many of the sources, conflicting reports, and the element of subjectivity in the ratings, but the author has evidently been fully cognizant of these.

On the whole, the work seems to be thoroughly done, it is well documented, the range of data is most extensive, the presentation is clear.

It is recommended to social psychologists as an excellent contribution to our knowledge of the position of the aged in primitive society.

HERBERT GURNEE.

Arizona State College at Tempe.

ORDAN, HARRY. *Social concepts and the child mind.* New York: King's Crown Press, 1945. Pp. 130.

Children continue to startle us by their familiarity with crime. Whether or not we place the blame on scare headlines, gangster movies, radio thrillers or local crime, the reviewer agrees with the author, that education should be used as a more powerful concept builder.

The present study was carried on in two Brooklyn schools, in areas differentiated with respect to the intellectual and socio-economic status of their school populations. A series of tests was constructed to measure the recognition of social problems by children; their ability to identify, to particularize and to discriminate social concepts.

The results indicate that intellectual maturity is of first importance in determining children's recognition of social problems; that socio-economic status makes little or no contribution; that recognition in the area

of crime, not emphasized in the school curriculum is found to be higher than recognition in other problem areas. The order of familiarity with concepts belonging to the particular areas was found to be the same for both school populations, namely crime, economics or government, health, war-peace and socio-ethical. On the basis of recognition, the order of the six areas was found to be the same on every grade level and for both school groups, namely, health, crime, economics or government, war-peace and socio-ethical.

As the author has pointed out, these findings may be accepted only tentatively due to limiting conditions met in defining and classifying social problem areas and concepts. Experimentation with widely different socio-economic areas would doubtless have yielded important data.

Effective instruction in social problems based on environmental and experiential changes, as evidenced by this and other similar studies, is the exception rather than the rule. This study should prove helpful to those responsible for curriculum development and text-book writing in this field.

RUTH F. BOLAND.

Lehigh University.

NOTICE

After November 1, 1946, all manuscripts, books for review, and correspondence concerning the *Psychological Bulletin* should be sent to the new Editor, PROFESSOR LYLE H. LANIER, *Department of Psychology, Vassar College, Poughkeepsie, New York.*

INDEX OF SUBJECTS

INDEX OF AUTHORS

ORIGINAL CONTRIBUTIONS, SHORT ARTICLES, SPECIAL REVIEWS, REPORTS, NOTES

BOOKS REVIEWED

BOOK REVIEWERS

Vol. 43, No. 6 November, 1946

Psychological Bulletin

EDITED BY

JOHN E. ANDERSON, UNIVERSITY OF MINNESOTA

WITH THE CO-OPERATION OF

S. H. BRITT, McCANN-ERICKSON, INC., NEW YORK; D. A. GRANT, UNIVERSITY OF WISCONSIN; W. T. HERON, UNIVERSITY OF MINNESOTA; W. A. HUNT, NORTHWESTERN UNIVERSITY; J. G. JENKINS, UNIVERSITY OF MARYLAND; D. G. MARQUIS, UNIVERSITY OF MICHIGAN; A. W. MELTON, UNIVERSITY OF MISSOURI; J. T. METCALF, UNIVERSITY OF VERMONT.

CONTENTS

PUBLISHED BI-MONTHLY BY

THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

1515 Massachusetts Ave., N.W., Washington 5, D.C.

Subscription price, $7.00 per year

Entered as second class mail matter at the post office at Washington, D.C., under the Act of March 3, 1879. Additional entry at the post office at Menasha, Wisconsin.

GEORGE BANTA PUBLISHING COMPANY, MENASHA, WISCONSIN